Certifying Equality With Limited Interaction*

Joshua Brody
Swarthmore College
joshua.e.brody@gmail.com

Amit Chakrabarti
Dartmouth College
amit.chakrabarti@dartmouth.edu

Ranganath Kondapally
Dartmouth College
rangak@cs.dartmouth.edu

David P. Woodruff
IBM Almaden
dpwoodru@us.ibm.com

Grigory Yaroslavtsev †
University of Pennsylvania
grigory@grigory.us

Abstract

The EQUALITY problem is usually one's first encounter with communication complexity and is one of the most fundamental problems in the field. Although its deterministic and randomized communication complexity were settled decades ago, we find several new things to say about the problem by focusing on three subtle aspects. The first is to consider the *expected* communication cost (at a worst-case input) for a protocol that uses limited interaction—i.e., a bounded number of rounds of communication—and whose error probability is zero or close to it. The second is to treat the *false negative* error rate separately from the *false positive* error rate. The third is to consider the *information cost* of such protocols. We obtain asymptotically optimal rounds-versus-cost tradeoffs for EQUALITY: both expected communication complexity and information complexity scale as $\Theta(\text{ilog}^{r-1} n)$, where r is the number of rounds and $\text{ilog}^k n = \log\log\cdots\log n$, with k logs. These bounds hold even when the false negative rate approaches 1. For the case of zero-error communication cost, we obtain essentially matching bounds, up to a tiny additive constant. We also provide some applications.

As an application of our information cost bounds, we obtain new bounded-round randomized lower bounds for the INTERSECTION problem, in which there are two players who hold subsets $S,T \subseteq [n]$. In many realistic scenarios, the sizes of S and T are significantly smaller than n, so we impose the constraint that $|S|,|T| \leq k$. We study the minimum number of bits the parties need to communicate in order to compute the entire intersection set $S \cap T$, using r rounds. We show that any r-round protocol has information cost (and thus communication cost) $\Omega(k\,\text{ilog}^r k)$ bits. We also give an $O(r)$-round protocol achieving $O(k\,\text{ilog}^r k)$ bits, which for $r = \log^* k$ gives a protocol with $O(k)$ bits of communication. This is in contrast to other basic problems such as computing the union or symmetric difference, for which $\Omega(k\log(n/k))$ bits of communication is required for any number of rounds.

*This is a full version containing results from papers "Certifying Equality With Limited Interaction" (RANDOM'14) and "Beyond Set Disjointness: The Communication Complexity of Finding the Intersection" (PODC'14) by the same authors.

†Part of this work was done while G.Y. was an intern at IBM Research, Almaden. G.Y. was also supported by the Warren Center Fellowship at the University of Pennsylvania and the Institute Postdoctoral Fellowship at Brown University, ICERM.

1 Introduction

1.1 Context

Over the last three decades, communication complexity [51] has proved itself to be among the most useful of abstractions in computer science. A number of basic problems in communication complexity have found a wide range of applications throughout the theory of computing, with EQUALITY, INDEX, and DISJOINTNESS being notable superstars.

Revisiting these basic problems and asking more nuanced questions or studying natural variants has extended their range of application. We highlight two examples. Our first example is DISJOINTNESS. The optimal $\Omega(n)$ lower bound for this problem [33, 48] was already considered one of the major results in communication complexity before DISJOINTNESS was revisited in the *multi-party* number-in-hand model to obtain a number of data stream lower bounds [3, 4, 15, 27] culminating in optimal space bounds for the (higher) frequency moments. Later, DISJOINTNESS was revisited in an *asymmetric* communication setting [46] yielding an impressive array of lower bounds for data structures in the cell-probe model. Very recently, DISJOINTNESS was revisited yet again in a *high-error* setting, yielding deep insights into extended formulations for the MAX-CLIQUE problem [9]. Our second example is INDEX. The straightforward $\Omega(n)$ lower bound on its one-way communication complexity [1] is already an important starting point for numerous other lower bounds. Revisiting INDEX in an *interactive* communication setting and considering communication tradeoffs has led to new classes of data stream lower bounds for memory-checking problems [39, 14, 16]. Separately, lower bounding the *quantum* communication complexity of INDEX [44] has led, among other things, to strong lower bounds for locally decodable codes [35, 18].

1.2 Our Results

In this work we revisit the EQUALITY problem: Alice and Bob each hold an n-bit string, and their task is to decide whether these strings are equal. This is arguably the most basic communication problem that admits a nontrivial protocol: using randomization and allowing a constant error rate, the problem can be solved with just $O(1)$ communication (this becomes $O(\log n)$ if one insists on private coins only); see, e.g., Kushilevitz and Nisan [37, Example 3.13] and Freivalds [26]. This is why a student's first encounter with communication complexity is usually through the EQUALITY problem. Such a fundamental problem deserves the most thorough of studies.

At first glance, the complexity of EQUALITY might appear "solved": its deterministic communication complexity is at least n, whereas its randomized complexity is $O(1)$ as noted above, as is its *information complexity* [6] (for more on this, see Section 1.3). However, one can ask the following more nuanced question. What happens if Alice and Bob want to be *certain* (or nearly certain) that their inputs are indeed equal when the protocol directs them to say so? And what happens if Alice and Bob want to run a protocol with limited interaction, i.e., a bounded number of back-and-forth rounds of communication?

Formally, let $\text{EQ}_n : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ denote the Boolean function that underlies this communication problem, defined by $\text{EQ}_n(x,y) = 1 \iff x = y$. Consider the zero-error case: the players must always correctly output $\text{EQ}_n(x,y)$ on every input (x,y). However, the players may use a randomized protocol and their goal is to minimize the *expected* number of bits they exchange. If their protocol is required to use only one round—this means that Alice sends a message to Bob, who then outputs the answer—then it is easy to see that Alice's message must uniquely identify her input to Bob. From this it is easy to show that on some input, x, Alice must send at least n bits to Bob, even in expectation.

Things improve a lot if one allows two rounds of communication—Alice sends a message to Bob, who replies to Alice, who then outputs the answer. Using standard techniques, Alice can send Bob a $\lceil \log n \rceil$-bit[1]

[1] Throughout this paper we use "log" to denote the logarithm to the base 2.

fingerprint of x. When $x \neq y$, this fingerprint fails to demonstrate that $\text{EQ}_n(x,y) = 0$ with probability at most $1/n$. If necessary, Bob responds to this failure by sending y to Alice, which costs only 1 bit in expectation. The net result is an expected communication cost of $O(\log n)$ on unequal inputs, and $O(n)$ on equal inputs. Generalizing this idea, we obtain an r-round protocol where the expected cost drops to $O(\text{ilog}^{r-1} n)$ on unequal inputs, where $\text{ilog}^j n := \log\log\cdots\log n$ (with j logs).

Our main high-level message in this work is that *the above tradeoff between the number of rounds and the communication cost is optimal*, and that this remains the case even allowing for some false positives, even allowing for a false negative rate of $1 - o(1)$, and even if we consider *information cost*. We shall get precise about information cost measures in Section 2, but for now we remark that an information cost lower bound is stronger than a communication cost bound, even in our expected-cost model.

While our main focus is on EQUALITY, our rounds-versus-information tradeoff can be applied to three other problems: OR-EQUALITY, DISJOINTNESS, and PRIVATE-INTERSECTION. (Based on developments since the initial announcement of our results [12], these derivative results can be proved, and in a sense strengthened, using alternative means: see the discussion at the end of Section 1.3. However, we feel there is value in our simpler and more direct approach.) It is well known that information cost has clean direct-sum properties [17, 4, 5]. Together with our results for EQUALITY, this easily gives us bounded-round randomized lower bounds for the OR-EQUALITY problem, whose underlying function is $\text{OREQ}_{n,k} : \{0,1\}^{nk} \times \{0,1\}^{nk} \to \{0,1\}$, defined by $\text{OREQ}_{n,k}(x_1,\ldots,x_k,y_1,\ldots,y_k) = \bigvee_{i=1}^{k} \text{EQ}_n(x_i,y_i)$: Alice holds each $x_i \in \{0,1\}^n$ and Bob holds each $y_i \in \{0,1\}^n$. Our lower bound is of the form $\Omega(k\,\text{ilog}^r k)$ for $k \ll 2^n$. It holds only for subconstant false positive rates (because EQ itself is too easy at a constant false positive rate); however the lower bound does apply under a false negative rate as high as $1 - o(1)$.

The OREQ problem is closely related to DISJOINTNESS, especially the variant called *small set disjointness* or k-DISJ_N. Here, Alice and Bob are given sets $A, B \subseteq [N]$ respectively[2], with the promise that $|A| \leq k$ and $|B| \leq k$, where $1 \leq k \leq N$. Their goal is to output 1 iff $A \cap B = \varnothing$. Using this close relation (see Lemma 7.3 for a formal treatment), we obtain bounded-round lower bounds for k-DISJ as well, also of the form $\Omega(k\,\text{ilog}^r k)$, provided $N \gg k^2$.

Yet another closely related problem is PRIVATE-INTERSECTION, which we also denote k-INT_N. Here, as in k-DISJ, Alice and Bob receive sets $A, B \subseteq [N]$ with $|A| \leq k$ and $|B| \leq k$. Each player should locally output the entire set $A \cap B$. The non-Boolean problem k-INT_N admits a similar $\Omega(k\,\text{ilog}^r k)$ lower bound, this time even in a constant-error setting. To complement this, we also give an upper bound of $O(k\,\text{ilog}^r k)$ for k-INT_N using $O(r)$ rounds; note that with $O(\log^* k)$ rounds this amounts to $O(k)$ communication. This should be contrasted with other basic problems such as computing the union or symmetric difference, for which $\Omega(k \log(n/k))$ bits of communication is required with any number of rounds. Given our protocol it is straightforward to obtain the same round/communication tradeoffs (up to an additive $O(\log k)$ in communication) for computing the *exact* Jaccard similarity $|A \cap B|/|A \cup B|$, the *exact* Hamming distance, the *exact* number of distinct elements, and the *exact* 1 and 2-rarity [21], all when $|A|, |B| \leq k$.

Our lower bound for k-INT_N applies directly to information cost, which is why we think of it as a lower bound for PRIVATE-INTERSECTION. A key property of information cost is that it is a measure of *privacy* of a protocol for a function f. Klauck [36] defines[3] the privacy of a protocol Π with respect to a distribution μ:

$$\text{PRIV}^{\mu}(\Pi) := \text{I}(X : \Pi(X,Y) \mid Y, f(X,Y)) + \text{I}(Y : \Pi(X,Y) \mid X, f(X,Y)).$$

This definition coincides with $\text{icost}^{\mu}(\Pi)$ up to the conditioning on $f(X,Y)$ in the mutual information expressions. However, in many cases, including this paper, this conditioning does not asymptotically affect the definition, and one has $\text{PRIV}^{\mu}(\Pi) = \Theta(\text{icost}^{\mu}(\Pi))$. One can then naturally define $\text{PRIV}_{\delta}(f) = \inf_{\delta\text{-error } \Pi} \max_{\text{input dist } \mu} \text{PRIV}^{\mu}(\Pi)$, and one has that $\text{PRIV}_{\delta}(f) = \Theta(\text{IC}_{\delta}(f))$.

[2] We use $[n]$ to denote the set $\{1,2,\ldots,n\}$.

[3] We have replaced the max in Klauck's definition with a sum; this agrees with Klauck's original definition up to a factor of 2.

There is a large body of work on solving EQUALITY privately. These are known as *private equality tests* in the cryptography and privacy literature [22, 43]. The harder problem PRIVATE-INTERSECTION is a fundamental problem studied in private datamining, see, e.g., Freedman et al. [25]. The problem is studied both under computational assumptions on the players, as in Freedman et al., and also using information-theoretic notions of privacy, such as $\mathrm{PRIV}_\delta(f)$, as in the work by Ada et al. [2]. In this context, it is worth noting that the bounded-round setting has a very good practical motivation: the number of rounds of a protocol may in fact influence its latency drastically while the actual number of bits communicated may not. This is because the more interactive protocols are, i.e., the larger the number of rounds, the more coordination is needed between the players, which may not be possible if, e.g., a player goes offline.

To obtain our information cost (hence, privacy) lower bound for PRIVATE-INTERSECTION, we combine our lower bound for EQ with a recent *direct sum theorem with aborts*, given by Molinaro et al. [42]. Roughly speaking, their theorem states that the information complexity of solving all k copies of a problem is k times the information complexity of solving each copy with a protocol that is allowed to output "abort" with a constant $1/10$ probability, but given that it does not output "abort" the protocol must be correct with a very high $1-1/k$ probability.[4] By changing such a protocol for EQUALITY so that whenever it would have output "abort", it instead declares that $x \neq y$, we show how to obtain an $\Omega(k \operatorname{ilog}^r k)$ information cost bound for k-INT_N for any r-round protocol with constant success probability. As $\mathrm{I}(\Pi : A \mid B, A \cap B) + \mathrm{I}(\Pi : B \mid A, A \cap B) = \mathrm{I}(\Pi : A \mid B) + \mathrm{I}(\Pi : B \mid A) \pm O(k)$, it follows that $\mathrm{PRIV}_{1/3}(\text{PRIVATE-INTERSECTION}) = \Omega(k \operatorname{ilog}^r k)$.

For a concise—yet technically precise—listing of our results, please see Section 2.

1.3 Related Work

The study of the EQUALITY problem dates back to the original communication complexity paper of Yao [51], who showed that the deterministic communication complexity of EQ_n is at least n, using a fooling set argument. Mehlhorn and Schmidt [40] developed the *rank lower bound* technique, which can recover this result. They further examined OR-EQUALITY, giving a lower bound of nk bits for deterministic protocols that compute $\mathrm{OREQ}_{n,k}$ via the rank technique. They also gave $O(n + \log n)$ and $O(n \log n)$ bounds for the nondeterministic and co-nondeterministic communication complexities of $\mathrm{OREQ}_{n,n}$, respectively. Furthermore, they studied the "Las Vegas" communication complexity of $\mathrm{OREQ}_{n,n}$, which brought them close to some of the things we study here. Specifically, they gave a zero-error private-coin randomized protocol such that the expected communication cost on any inputs $(x_1, \ldots, x_n, y_1, \ldots, y_n)$ is at most $O(n(\log n)^2)$.

Feder et al. [23] studied the randomized communication complexity of EQUALITY in the direct-sum setting. Here, players have k strings each and must compute $(\mathrm{EQ}_n(x_1, y_1), \ldots, \mathrm{EQ}_n(x_k, y_k))$: thus, the output is a k-bit string. Feder et al. showed that $O(k)$ communication suffices to compute EQUALITY on all k instances, with error *exponentially* small in k. This shows that the "amortized" communication complexity of EQ_n is $O(1)$, even under tiny error. More recently, Braverman and Rao [10] showed that amortized communication complexity nearly equals *information* complexity. Furthermore, Braverman [6] gave a specific protocol for EQ_n that has zero error and achieves internal information cost $O(1)$ regardless of the input distribution.

The problem $\mathrm{OREQ}_{n,k}$ is potentially easier than the k-fold direct sum of EQ_n, and has itself been studied a few times before. Chakrabarti et al. [17] showed that its simultaneous-message complexity is $\Omega(k\sqrt{n})$,

[4] It is crucial for us to use a strong direct sum theorem of [42] in the lower bound for PRIVATE-INTERSECTION. Unlike generic direct sum and direct product theorems which apply to any function, the strong direct sum of [42] only applies to EQUALITY-type functions but gives a much stronger guarantee in the constant error regime that we study here. This is in contrast with the bounded round direct product theorem of [30, 31] (and other similar results such as [32]), who show that for r-round public-coin randomized information complexity $\mathrm{IC}^{r,\mathrm{pub}}_{1-(1-\varepsilon/2)^{\Omega(k\varepsilon^2/r^2)}}(f^k) = \Omega\big(\ (\varepsilon k/r) \cdot (IC^{r,\mathrm{pub}}_\varepsilon(f) - O(r^2/\varepsilon^2))\ \big)$, where $\varepsilon > 0$ is arbitrary (the results of [30, 31] are stated in terms of communication complexity but their techniques also imply an information cost lower bound). One cannot apply this theorem to our problem, as one would need to set $\varepsilon = \Theta(k^{-1/3})$ to obtain our results. Because $\mathrm{IC}^{r,\mathrm{pub}}_{1/k^{1/3}}(\text{EQUALITY}) = o(k^{2/3})$ this theorem gives a trivial bound.

which is k times the complexity of EQ_n in that model. More recently, Kushilevitz and Weinreb [38] studied the deterministic complexity of $\text{OREQ}_{n,k}$ under the promise that $x_i = y_i$ for at most one $i \in [k]$. Computing $\text{OREQ}_{n,k}$ under this "$0/1$ intersection" promise is closely related to the clique-vs.-independent set problem. In this problem, Alice is given a clique in a graph. Bob is given an independent set, and they must decide if their inputs intersect. Kushilevitz and Weinreb were able to show that computing $\text{OREQ}_{n,k}$ under this promise still requires $\Omega(kn)$ communication whenever $k \leq n/\log n$. Extending this lower bound to the setting where $k = n$ is an important open problem, with several implications.

For the k-DISJ problem, Håstad and Wigderson [29] gave an $O(k)$-bit randomized protocol; a matching lower bound follows easily from the $\Omega(n)$ lower bound for general DISJOINTNESS. The Håstad–Wigderson protocol is clever and crucially exploits both the public randomness and the interactive communication between players. Sağlam and Tardos [49] extend this protocol to interpolate between the one-round and unbounded-round situations, showing that to compute k-DISJ in r rounds, $\Theta(k \operatorname{ilog}^r k)$ bits are necessary and sufficient. This lower bound extends tight $\Omega(k \log k)$ lower bounds for one-round protocols recently given by Dasgupta, Kumar, and Sivakumar [20] and by Buhrman et al. [13].

A different thread of research has been studying the relationship between information and communication complexities in the abstract, i.e., for general functions and relations. Most results in this thread have been *protocol compression* results [28, 5, 10, 31, 11] that show that information-efficient protocols can be turned into communication-efficient ones. Therefore, to some extent, they imply information cost lower bounds based on communication cost lower bounds. However, due to the the subtleties of our error parametrization, we cannot directly infer our information complexity lower bounds from communication lower bounds plus existing compression results. For instance, the communication lower bounds for OR-EQUALITY and DISJOINTNESS due to Sağlam and Tardos [49]—which we learned of following the initial announcement of this work [12]—imply lower bounds for information complexity of those two problems when combined with compression results of Harsha et al. [28]. Additionally, with the recent direct product theorem for bounded-round communication complexity of Jain et al. [31] and the existing result equating information and amortized communication of Braverman and Rao [10], these results also extend to give information complexity lower bounds for bounded-round protocols for EQUALITY. Still, EQUALITY is one of the most important communication complexity problems; as such, it deserves careful study. Our information cost lower bounds are more direct and shed more light on this important problem. In particular, previous results do not differentiate between errors for false positives and false negatives and therefore cannot admit the high false negative rate our bounds apply to.

The recent work of Braverman et al. [8] is similar in spirit to some of our results. They consider zero-error communication protocols for the even more fundamental AND function, obtaining exact information cost bounds. From this they derive nearly exact communication bounds for low-error protocols for DISJOINTNESS and k-DISJ. They also consider rounds-vs.-information tradeoffs for AND, showing that the information complexity of r-round protocols decays as $\Theta(1/r^2)$. Our work shows that the information complexity of EQUALITY decays exponentially with each additional round.

1.4 Road Map

The rest of the paper is organized as follows. Section 2 gives careful definitions of our model of computation and error and cost measures, followed by a listing of all our results. The listing provides pointers to later sections of the paper where these results are proved. Section 3 gives basic definitions and lemmas relating to information theory.

The next two sections provide some warm-up. Section 4 gives upper bounds for EQUALITY including the iterated-log upper bound described informally above. Section 5 gives matching lower bounds for expected communication complexity, first under zero error and then under two-sided error. Though the proofs in Sections 4 and 5 are not too complex, the combined story they tell is important. Together, these results

paint a nearly complete picture of the behavior of EQUALITY in a bounded-round expected-communication setting, and highlight the importance of studying YES and NO instances separately. Nevertheless, the reader who is interested solely in information cost lower bounds may safely skip these sections.

Section 6 contains the full proof of our Main Theorem, which gives an information cost lower bound for EQUALITY. Section 7 obtains lower bounds for OREQ and k-DISJ as quick applications of the Main Theorem, and a lower bound for PRIVATE-INTERSECTION after suitably extending the Main Theorem to the setting of protocols with abortion. Finally, Sections 8 and 9 give our protocols for k-INT$_N$ in the two-party and multi-party settings, respectively.

2 Definitions and Formal Statement of Results

Throughout this paper we reserve the symbols "n" for input length of EQUALITY instances, "k" for list length of OR-EQUALITY instances and set size of k-DISJ instances, and "N" for universe size of k-DISJ or k-INT$_N$ instances. Many definitions and results will be parametrized by these quantities but we shall not always make this parametrization explicit. We tacitly assume that n, k and N are sufficiently large integers.

Unless otherwise stated, all communication protocols appearing in this paper are public-coin randomized protocols involving two players named Alice and Bob. Because our work concerns expected communication cost in a bounded-round setting, we make the following careful definition of what communication is allowed. In each round, the player whose turn it is to speak sends the other player a message from a *prefix-free* subset[5] of $\{0,1\}^*$. This subset can depend on the communication history. After the final round in the protocol, the player that receives the last message announces the output: this announcement does not count as a round.

Let $\mathscr{P}$ be a communication protocol that takes inputs $(x,y) \in \mathscr{X} \times \mathscr{Y}$. The *transcript* of $\mathscr{P}$ on input (x,y) is defined to be the concatenation of the messages sent by the players, in order, as they execute $\mathscr{P}$ on (x,y). We denote this transcript by $\mathscr{P}(x,y)$ and remark that it is, in general, a random variable. We include the output as the final "message" in the transcript. We denote the output of a transcript $\mathfrak{t}$ by $\mathrm{out}(\mathfrak{t})$. We denote the length of a binary string z by $|z|$. The *communication cost* and *worst-case communication cost* of $\mathscr{P}$ on input (x,y) are defined to be

$$\mathrm{cost}(\mathscr{P};x,y) := \mathbb{E}\left[|\mathscr{P}(x,y)|\right], \quad \text{and} \quad \mathrm{cost}^*(\mathscr{P};x,y) := \max|\mathscr{P}(x,y)|,$$

where the expectation and the max are taken over the protocol's random coin tosses.

We now define complexity measures based on this notion of communication cost. Ordinarily we would just define the communication complexity of a function f as the minimum over protocols for f of the worst-case (over all inputs) cost of the protocol. When $f = \mathrm{EQ}_n$, such a measure turns out to be too punishing, and hides the subtleties that we seek to study. Notice that the r-round protocol outlined in Section 1.2 achieves its cost savings only on unequal inputs, i.e., on $f^{-1}(0)$. On inputs in $f^{-1}(1)$, the protocol ends up costing at least n bits. The intuition is that it is much cheaper for Alice and Bob to *refute* the purported equality of their inputs than to *verify* it. Indeed, verification is so hard that interaction has no effect on the verification cost, whereas each additional round of communication decreases refutation cost exponentially.

In fact, this intuition can be turned into precise theorems, both in zero-error and positive-error settings, as we shall see. To formalize things, we now define a family of complexity measures.

Definition 2.1 (Cost, Error, and Complexity Measures)**.** Let $\mathscr{P}$ be a protocol that computes a Boolean function $f : \mathscr{X} \times \mathscr{Y} \to \{0,1\}$. We define its *refutation cost*, *verification cost*, *overall cost*, *refutation error* (or false positive rate, or soundness error), and *verification error* (or false negative rate, or completeness

[5] A set of strings is said to be prefix-free if no string in the set is a proper prefix of any other.

error) as follows, respectively:

$$\begin{aligned}
\mathrm{rcost}(\mathscr{P}) &:= \max_{(x,y)\in f^{-1}(0)} \mathrm{cost}(\mathscr{P};x,y)\,,\\
\mathrm{vcost}(\mathscr{P}) &:= \max_{(x,y)\in f^{-1}(1)} \mathrm{cost}(\mathscr{P};x,y)\,,\\
\mathrm{cost}(\mathscr{P}) &:= \max_{(x,y)\in \mathscr{X}\times\mathscr{Y}} \mathrm{cost}(\mathscr{P};x,y)\,,\\
\mathrm{rerr}(\mathscr{P}) &:= \max_{(x,y)\in f^{-1}(0)} \Pr[\mathrm{out}(\mathscr{P}(x,y))=1]\,,\\
\mathrm{verr}(\mathscr{P}) &:= \max_{(x,y)\in f^{-1}(1)} \Pr[\mathrm{out}(\mathscr{P}(x,y))=0]\,.
\end{aligned}$$

Let λ be a probability distribution on the input space $\mathscr{X}\times\mathscr{Y}$. We then define the λ-distributional error $\mathrm{err}^{\lambda}(\mathscr{P})$ as well as the λ-distributional refutation cost, etc., as follows:

$$\begin{aligned}
\mathrm{rcost}^{\lambda}(\mathscr{P}) &:= \mathbb{E}_{(X,Y)\sim\lambda}[\mathrm{cost}(\mathscr{P};X,Y)\mid f(X,Y)=0]\,,\\
\mathrm{vcost}^{\lambda}(\mathscr{P}) &:= \mathbb{E}_{(X,Y)\sim\lambda}[\mathrm{cost}(\mathscr{P};X,Y)\mid f(X,Y)=1]\,,\\
\mathrm{cost}^{\lambda}(\mathscr{P}) &:= \mathbb{E}_{(X,Y)\sim\lambda}[\mathrm{cost}(\mathscr{P};X,Y)]\,,\\
\mathrm{rerr}^{\lambda}(\mathscr{P}) &:= \mathbb{E}_{(X,Y)\sim\lambda}[\Pr[\mathrm{out}(\mathscr{P}(X,Y))=1]\mid f(X,Y)=0]\,,\\
\mathrm{verr}^{\lambda}(\mathscr{P}) &:= \mathbb{E}_{(X,Y)\sim\lambda}[\Pr[\mathrm{out}(\mathscr{P}(X,Y))=0]\mid f(X,Y)=1]\,,\\
\mathrm{err}^{\lambda}(\mathscr{P}) &:= \mathbb{E}_{(X,Y)\sim\lambda}[\Pr[\mathrm{out}(\mathscr{P}(X,Y))\neq f(X,Y)]\,.
\end{aligned}$$

We shall usually restrict $\mathscr{P}$ to be deterministic when considering these distributional measures. Although these measures depend on both $\mathscr{P}$ and f, we do not indicate f in our notation to keep things simple.

Let $r\geq 1$ be an integer and let $\varepsilon,\delta\in[0,1]$ be reals. We define the r-round randomized *refutation complexity* and r-round λ-distributional refutation complexity of f as follows, respectively:

$$\begin{aligned}
\mathrm{R}^{(r),\mathrm{ref}}_{\varepsilon,\delta}(f) &:= \min\{\mathrm{rcost}(\mathscr{P}):\ \mathscr{P}\text{ uses }r\text{ rounds},\mathrm{rerr}(\mathscr{P})\leq\varepsilon,\mathrm{verr}(\mathscr{P})\leq\delta\}\,,\\
\mathrm{D}^{\lambda,(r),\mathrm{ref}}_{\varepsilon,\delta}(f) &:= \min\{\mathrm{rcost}^{\lambda}(\mathscr{P}):\ \mathscr{P}\text{ is deterministic and uses }r\text{ rounds},\mathrm{rerr}^{\lambda}(\mathscr{P})\leq\varepsilon,\mathrm{verr}^{\lambda}(\mathscr{P})\leq\delta\}\,.
\end{aligned}$$

We also define measures of *verification complexity* and *overall complexity* analogously, replacing "rcost" above with "vcost" and "cost" respectively, and denote them by

$$\mathrm{R}^{(r),\mathrm{ver}}_{\varepsilon,\delta}(f),\ \mathrm{D}^{\lambda,(r),\mathrm{ver}}_{\varepsilon,\delta}(f),\ \mathrm{R}^{(r)}_{\varepsilon,\delta}(f),\text{ and }\mathrm{D}^{\lambda,(r)}_{\varepsilon,\delta}(f)\,,$$

respectively. We define the *total complexity* of f as follows:

$$\begin{aligned}
\mathrm{R}^{*,(r)}_{\varepsilon,\delta}(f) &:= \min\{\mathrm{cost}^{*}(\mathscr{P}):\ \mathscr{P}\text{ uses }r\text{ rounds},\mathrm{rerr}(\mathscr{P})\leq\varepsilon,\mathrm{verr}(\mathscr{P})\leq\delta\}\,,\text{ where}\\
\mathrm{cost}^{*}(\mathscr{P}) &:= \max_{(x,y)\in\mathscr{X}\times\mathscr{Y}}\mathrm{cost}^{*}(\mathscr{P};x,y)\,.
\end{aligned}$$

Notice that refutation, verification, and overall complexities use (expected) communication cost as the underlying measure, whereas total complexity uses the (more standard) worst-case communication cost.

Definition 2.2 (Information Cost and Complexity)**.** Let $\mathscr{P},f$, and λ be as above, and suppose the players in $\mathscr{P}$ are allowed to use private coins in addition to a public random string $\mathfrak{R}$. The λ-*information cost* of $\mathscr{P}$ and the r-round λ-*information complexity* of f are defined as follows, respectively:

$$\begin{aligned}
\mathrm{icost}^{\lambda}(\mathscr{P}) &:= \mathrm{I}(XY:\mathscr{P}(X,Y)\mid\mathfrak{R})\,,\\
\mathrm{IC}^{\lambda,(r)}_{\varepsilon,\delta}(f) &:= \inf\{\mathrm{icost}^{\lambda}(\mathscr{P}):\ \mathscr{P}\text{ uses }r\text{ rounds},\mathrm{rerr}(\mathscr{P})\leq\varepsilon,\mathrm{verr}(\mathscr{P})\leq\delta\}\,.
\end{aligned}$$

where $\mathrm{I}(\_:\_\mid\_)$ denotes conditional mutual information. For readers familiar with recent literature on information complexity [5, 6], we note that this is technically the "external" information cost rather than the "internal" one. However, we shall study information costs mostly with respect to a uniform input distribution, and in this setting there is no difference between external and internal information cost [10].

It has long been known that information complexity lower bounds standard worst-case communication complexity: this was the main reason for defining the notion [17]. The simple proof boils down to

$$\mathrm{I}(XY : \mathscr{P}(X,Y) \mid \mathfrak{R}) \leq \mathrm{H}(\mathscr{P}(X,Y)) \leq \max|\mathscr{P}(X,Y)|\,.$$

In our setting, with communication cost defined in the expected sense, it is still the case that

$$\mathrm{IC}^{\lambda,(r)}_{\varepsilon,\delta}(f) \leq \mathrm{R}^{(r)}_{\varepsilon,\delta}(f) \tag{1}$$

This time the proof boils down to the inequality $\mathrm{H}(\mathscr{P}(X,Y)) \leq \mathbb{E}[|\mathscr{P}(X,Y)|]$, which follows from Shannon's source coding theorem (see Fact 3.6 below).

2.1 Summary of Results for Equality

The functions EQ_n and $\mathrm{OREQ}_{n,k}$ have been defined in Section 1 already. To formalize our bounds for these problems, we introduce the iterated logarithm functions $\mathrm{ilog}^k : \mathbb{R}_+ \to \mathbb{R}_+$, which are defined as follows.

$$\begin{aligned}
&\mathrm{ilog}^0 z := \max\{1, z\}\,, \quad \forall z \in \mathbb{R}_+\,,\\
&\mathrm{ilog}^k z := \max\{1, \log(\mathrm{ilog}^{k-1} z)\}\,, \quad \forall k \in \mathbb{N}, z \in \mathbb{R}_+\,.
\end{aligned}$$

For all practical purposes, we may pretend that $\mathrm{ilog}^0 = \mathrm{id}$, and $\mathrm{ilog}^k = \log \circ \mathrm{ilog}^{k-1}$, for $k \in \mathbb{N}$.

We use ξ to denote the uniform distribution on $\{0,1\}^n$, and put $\mu_u := \xi \otimes \xi$. Thus μ_u is the uniform distribution on inputs to EQ_n. Strictly speaking these should be denoted ξ_n and $\mu_{u,n}$, but we choose to let n be understood from the context. In all our complexity bounds, we tacitly assume that n is sufficiently large. The various parts of the summary theorems below are proved later in the paper, and we indicate on the right where these detailed proofs can be found.

Theorem 2.3 (Zero-Error Bounds)**.** *The complexity of* EQUALITY *satisfies the following bounds:*

1. $\mathrm{R}^{(r),\mathrm{ref}}_{0,0}(\mathrm{EQ}_n) \leq \mathrm{ilog}^{r-1} n + 3.$
2. $\mathrm{R}^{(r),\mathrm{ver}}_{0,0}(\mathrm{EQ}_n) \leq n.$
3. $\mathrm{R}^{(r),\mathrm{ref}}_{0,0}(\mathrm{EQ}_n) = \mathrm{D}^{\mu_u,(r),\mathrm{ref}}_{0,0}(\mathrm{EQ}_n) \geq \mathrm{ilog}^{r-1} n - 1.$ *[Theorem 5.2]*
4. $\mathrm{R}^{(r),\mathrm{ver}}_{0,0}(\mathrm{EQ}_n) = \mathrm{D}^{\mu_u,(r),\mathrm{ver}}_{0,0}(\mathrm{EQ}_n) \geq n.$ *[Theorem 5.5]*

Notice that these bounds are almost completely tight, differing at most by the tiny additive constant 4. Next, we allow our protocols some error. We continue to have very tight bounds for the verification cost (the case of one-sided error is especially interesting: just set $\delta = 0$ in the results below), and we have asymptotically tight bounds in the other cases. To better appreciate the next several bounds, let us first consider the "trivial" one-round protocol for EQ_n that achieves ε refutation error. This protocol communicates $\min\{n, \log(1/\varepsilon)\}$ bits: it's as though the instance size drops from n to $\min\{n, \log(1/\varepsilon)\}$ when we allow this refutation error. This motivates the following definition.

Definition 2.4 (Effective Instance Size)**.** When considering protocols for EQ_n with refutation and verification errors bounded by ε and δ, respectively, we define the effective instance size to be

$$\hat{n} := \min\{n + \log(1-\delta), \log((1-\delta)^2/\varepsilon)\}\,.$$

Theorem 2.5 (Two-Sided-Error Bounds)**.** *The complexity of* EQUALITY *satisfies the following bounds:*

5. $\mathrm{R}^{(r),\mathrm{ref}}_{\varepsilon,\delta}(\mathrm{EQ}_n) \leq (1-\delta)\,\mathrm{ilog}^{r-1} \hat{n} + 5.$ *[Corollary 4.4]*

6. $\mathrm{R}^{(r),\mathrm{ver}}_{\varepsilon,\delta}(\mathrm{EQ}_n) \leq (1-\delta)\hat{n}+3$. *[Corollary 4.5]*

7. $\mathrm{D}^{\mu_u,(r),\mathrm{ver}}_{\varepsilon,\delta}(\mathrm{EQ}_n) \geq (1-\delta)(\hat{n}-1)$. *[Theorem 5.13]*

8. $\mathrm{R}^{(r),\mathrm{ver}}_{\varepsilon,\delta}(\mathrm{EQ}_n) \geq \frac{1}{8}(1-\delta)^2(\hat{n}+\log(1-\delta)-5)$. *[Theorem 5.14]*

9. $\mathrm{D}^{\mu_u,(r),\mathrm{ref}}_{\varepsilon,\delta}(\mathrm{EQ}_n) = \Omega((1-\delta)^2 \operatorname{ilog}^{r-1}\hat{n})$. *This bound holds for all ε,δ such that $\delta \leq 1-2^{-n/2}$ and $\varepsilon/(1-\delta)^2 < 1/8$.* *[Theorem 5.11]*

10. $\mathrm{R}^{(r),\mathrm{ref}}_{\varepsilon,\delta}(\mathrm{EQ}_n) = \Omega((1-\delta)^3 \operatorname{ilog}^{r-1}\hat{n})$. *This bound holds for all ε,δ such that $\delta \leq 1-2^{-n/2}$ and $\varepsilon/(1-\delta)^3 \leq 1/64$.* *[Theorem 5.12]*

Observe that the "constant refutation error" setting $\varepsilon = O(1)$ is not very interesting, as it makes these complexities constant. But observe also that the situation is very different for the verification error, δ: we continue to obtain strong lower bounds even when δ is very close to 1. This is in accordance with our intuition that verification (of equality) is much harder than refutation.

Finally, we turn to information complexity and arrive at the most important result of this paper. For readers curious about the implications of protocol compression results for the information complexity of EQUALITY, we refer the reader to the discussion in Section 1.3.

Theorem 2.6 (Main Theorem: Information Complexity Bound)**.** *Suppose $\delta \leq 1-8(\operatorname{ilog}^{r-2}\hat{n})^{-1/8}$. Then*

11. $\mathrm{IC}^{\mu_u,(r)}_{\varepsilon,\delta}(\mathrm{EQ}_n) = \Omega((1-\delta)^3 \operatorname{ilog}^{r-1}\hat{n})$. *[Theorem 6.7]*

2.2 On Yao's Minimax Lemma

Distributional lower bounds imply worst-case randomized ones by an averaging argument that constitutes the "easy" direction of Yao's minimax lemma [50]. Yet, in Theorem 2.5 we claim somewhat weaker randomized bounds than the corresponding distributional ones. The reason is that in our setting, the averaging argument will need to fix the random coins of a protocol so as to preserve multiple measures (e.g., refutation error as well as cost). Though this is easily accomplished, we pay a penalty of small constant factor increase in our measures.

Ironically, the "hard" direction of Yao's minimax lemma is particularly easy in the case of EQ_n, because EQUALITY is in a sense *uniform self-reducible*. See Theorem 4.3, where we show how to turn a protocol designed for the uniform distribution into a randomized one with worst-case guarantees. In this way, the uniform distribution is provably the hardest distribution for EQUALITY.

2.3 Applications of the Main Theorem

Our main theorem can be used to prove the following lower bounds for OR-EQUALITY, DISJOINTNESS, and PRIVATE-INTERSECTION. We now summarize our results for the functions $\mathrm{OREQ}_{n,k}$, k-DISJ_N and k-INT_N, which were defined in Section 1. Whenever δ appears in these results, it needs to be bounded sufficiently away from 1. Similarly, ε needs to be nonnegative, and n and N need to be sufficiently large. We state these results more precisely in Section 7.

Theorem 2.7. *The following lower bounds hold:*

1. *For $k \leq 2^{0.99n}$, we have* $\mathrm{R}^{(r)}_{k^{-1},0.99}(\mathrm{OREQ}_{n,k}) = \Omega(k \operatorname{ilog}^r k)$. *[Theorem 7.1]*

2. *For $N \geq k^{2.01}$, we have* $\mathrm{R}^{(r)}_{0.99,\ k^{-1}}(k\text{-}\mathrm{DISJ}_N) = \Omega(k \operatorname{ilog}^r k)$. *[Theorem 7.5]*

3. *For $N \geq k^{2.01}$, we have* $\mathrm{R}^{(r)}_{1/3,1/3}(k\text{-}\mathrm{INT}_N) = \Omega(k \operatorname{ilog}^r k)$. *[Theorem 7.8]*

Based on developments since the original announcement of our main theorem [12], one can in fact strengthen the first two results above. Sağlam and Tardos give $\Omega(k \operatorname{ilog}^r k)$ lower bounds on $\mathrm{R}^{(r)}_{1/3,1/3}(\mathrm{OREQ}_{n,k})$ and $\mathrm{R}^{(r)}_{1/3,1/3}(k\text{-DISJ}_N)$; the improvement lies in not requiring subconstant error. The third result above can now be derived in another way. A communication lower bound on PRIVATE-INTERSECTION follows directly from that on k-DISJ. One can then use the optimal bounded-round protocol compression result of [28] (Lemma V.3) to derive the information cost lower bound.

We remind the reader that our main objective in this paper is the thorough study of EQUALITY, including the direct development of information cost bounds for bounded-round protocols and the analysis of verification vs. refutation error.

2.4 Upper Bound Results for Set Intersection

In Section 8 we give a randomized protocol for k-INT_N which achieves the optimal $O(k)$ bits of communication, and simultaneously achieves $O(\log^* k)$ number of rounds. Our number of rounds provides a significant improvement on the earlier $O(\log k)$ rounds needed to achieve the optimal $O(k)$ bits of communication given in previous work for set disjointness [29]. We also provide a more refined tradeoff, showing that with $O(r)$ rounds, one can achieve communication $O(k \operatorname{ilog}^r k)$.

Theorem 2.8. *For every integer $r > 0$ there exists a $6r$-round constructive public-coin protocol for k-INT_N with total expected communication $O(k \operatorname{ilog}^r k)$ and success probability $1 - 1/\operatorname{poly}(k)$.*

In Section 9 we extend this to the setting in which there are m players in the private messages model [7, 47] and give a protocol with $O(k \operatorname{ilog}^r k)$ average communication per player, expected number of rounds $O(r \cdot \max(1, \frac{\log m}{k}))$, and error probability $1 - 1/2^k$. We give a similar guarantee with a worst-case communication bound per player.

Our protocols for two players are communication-optimal, up to a constant factor in the number of rounds r, in light of the results above. For m players and $O(\log^* k \cdot \max(1, \frac{\log m}{k}))$ rounds, our $O(mk)$ communication is also optimal up to constant factors [7, 47].

3 Preliminaries

Here we collect some basic facts from probability theory and information theory. Then we outline the theory of *protocols with abortion*, which is used in the final sections of the paper while studying direct sum questions.

3.1 Probability, Entropy and Mutual Information

We will use the following fact about collision probability of a random function.

Fact 3.1. Given a subset $S \subseteq [n]$ for size $|S| \geq 2$, $i \geq 0$ and $t = \Theta(|S|^{i+2})$, a random function $h\colon [n] \to [t]$ has no collisions with probability at least $1 - 1/|S|^i$, namely for all $x, y \in S$ such that $x \neq y$ it holds that $h(x) \neq h(y)$. Moreover, a random hash function satisfying such a guarantee can be constructed using only $O(\log n)$ random bits.

Definition 3.2. Let λ be a probability distribution on a finite set S and let $T \subseteq S$ be an event with $\lambda(T) \neq 0$. We write $\lambda \mid T$ to denote the distribution obtained by conditioning λ on T. To be explicit, $\lambda \mid T$ is given by

$$(\lambda \mid T)(x) = \begin{cases} 0, & \text{if } x \notin T, \\ \lambda(x)/\lambda(T), & \text{if } x \in T. \end{cases}$$

Also, we write $\mathrm{H}(\lambda)$ to denote the entropy of a random variable distributed according to λ, i.e., $\mathrm{H}(\lambda) = \mathrm{H}(X)$, where $X \sim \lambda$.

Lemma 3.3. *With λ, S and T as above, let $f : S \to \mathbb{R}_+$ be a nonnegative function. Then $\mathbb{E}_{X\sim\lambda|T}[f(X)] \leq \mathbb{E}_{X\sim\lambda}[f(X)]/\lambda(T)$.*

We collect together some basic results in information theory whose proofs can be found in any standard textbook, e.g., Cover and Thomas [19, Chapter 2].

Fact 3.4. Let X, Y, Z and $X_1, X_2, \ldots$ denote random variables, possibly correlated. Let $\mathrm{supp}(X)$ denote the support set of X. The following facts hold.

1. Entropy span: $0 \leq \mathrm{H}(X) \leq \log|\mathrm{supp}(X)|$.
2. $\mathrm{H}(X \mid Y) \leq \mathrm{H}(X)$, and thus $\mathrm{I}(X : Y) \geq 0$.
3. Chain rule: $\mathrm{I}(X_1, X_2, \ldots, X_n : Y \mid Z) = \sum_{i=1}^{n} \mathrm{I}(X_i : Y \mid X_1, \ldots, X_{i-1}, Z)$.
4. Subadditivity: $\mathrm{H}(X, Y \mid Z) \leq \mathrm{H}(X \mid Z) + \mathrm{H}(Y \mid Z)$, where the equality holds if and only if X and Y are independent conditioned on Z.
5. Fano's inequality: Let A be a random variable, which can be used as "predictor" of X, namely there exists a function g such that $\Pr[g(A) = X] \geq 1 - \delta$ for some $\delta < 1/2$. If $|\mathrm{supp}(X)| \geq 2$ then

$$\mathrm{H}(X \mid A) \leq \delta \log(|\mathrm{supp}(X)| - 1) + \mathrm{H}_b(\delta),$$

where $\mathrm{H}_b(\delta) = \delta \log(1/\delta) + (1-\delta)\log(1/(1-\delta))$ is the binary entropy function.

Fact 3.5 (Kraft Inequality)**.** Let $S \subseteq \{0,1\}^*$ be a prefix-free set. Then

$$\sum_{x \in S} 2^{-|x|} \leq 1.$$

Fact 3.6 (Source Coding Theorem)**.** Let X be a random variable taking values in a prefix-free set $S \subseteq \{0,1\}^*$. Then

$$\mathbb{E}[\,|X|\,] \geq \mathrm{H}(X).$$

Lemma 3.7. *Let Z, W be jointly distributed random variables. Let $\mathscr{E}$ be an event. Then,*

$$\mathrm{I}(Z : W) \geq \Pr[\mathscr{E}]\, \mathrm{I}(Z : W \mid \mathscr{E}) - 1.$$

Proof. Let D be the indicator random variable for $\mathscr{E}$. Then we have

$$\mathrm{I}(Z : W \mid D) = \Pr[\mathscr{E}]\, \mathrm{I}(Z : W \mid \mathscr{E}) + \Pr[\neg\mathscr{E}]\, \mathrm{I}(Z : W \mid \neg\mathscr{E}) \geq \Pr[\mathscr{E}]\, \mathrm{I}(Z : W \mid \mathscr{E}). \tag{2}$$

Note that $\mathrm{I}(Z : D \mid W) \leq \mathrm{H}(D \mid W) \leq \mathrm{H}(D) \leq 1$. Using the chain rule for mutual information twice, we get

$$\mathrm{I}(Z : W \mid D) \leq \mathrm{I}(Z : WD) = \mathrm{I}(Z : W) + \mathrm{I}(Z : D \mid W) \leq \mathrm{I}(Z : W) + 1. \tag{3}$$

The lemma follows by combining inequalities (2) and (3). □

To appreciate the next two lemmas, it will help to imagine that $d \ll n$.

Lemma 3.8. *Let Z, W be jointly distributed random variables, with Z taking values in $\{0,1\}^n$, and let $\mathscr{E}$ be an event. Then*

$$\mathrm{H}(Z \mid W) \geq n - d \implies \mathrm{H}(Z \mid W, \mathscr{E}) \geq n - (d+1)/\Pr[\mathscr{E}].$$

In particular, taking W to be a constant, we have $\mathrm{H}(Z) \geq n - d \implies \mathrm{H}(Z \mid \mathscr{E}) \geq n - (d+1)/\Pr[\mathscr{E}]$.

Proof. We use the fact that the entropy of Z can be at most n, even after arbitrary conditioning. This gives

$$\begin{aligned} n - d &\leq \mathrm{H}(Z \mid W) \\ &= \Pr[\mathscr{E}]\,\mathrm{H}(Z \mid W, \mathscr{E}) + (1 - \Pr[\mathscr{E}])\,\mathrm{H}(Z \mid W, \neg\mathscr{E}) + \mathrm{H}_b(\Pr[\mathscr{E}]) \\ &\leq \Pr[\mathscr{E}]\,\mathrm{H}(Z \mid W, \mathscr{E}) + (1 - \Pr[\mathscr{E}])n + 1\,. \end{aligned}$$

The lemma follows by rearranging the above inequality. □

Lemma 3.9. *Let Z be a random variable taking values in $\{0,1\}^n$ and let $z \in \{0,1\}^n$. Then*

$$\mathrm{H}(Z) \geq n - d \implies \Pr[Z = z] \leq (d+1)/n\,.$$

Proof. The lemma follows by rearranging the following inequality, which is a consequence of Lemma 3.8:

$$0 = \mathrm{H}(Z \mid Z = z) \geq n - \frac{d+1}{\Pr[Z = z]}\,.$$ □

3.2 Protocols with Abortion

For our eventual lower bound on PRIVATE-INTERSECTION (Section 7.3), we shall need the concept of communication protocols that are allowed to *abort*. Consider a communication problem given by a function $f : \mathscr{X} \times \mathscr{Y} \to \mathscr{Z}$, and a protocol $\mathscr{P}$ that attempts to compute f. We shall allow $\mathscr{P}$ to output the special value $\perp$, indicating "abort," in addition to values in $\mathscr{Z}$. The next definition captures the desired semantics of such a protocol.

Definition 3.10 (Protocols with abortion)**.** Let f be a function and $\mathscr{P}_D$ a *deterministic* protocol of the above form, and let μ be a probability distribution over $\mathscr{X} \times \mathscr{Y}$, the domain of f. We say that $\mathscr{P}_D$ (β, δ)-computes f with respect to μ if, with $(X,Y) \sim \mu$, we have

1. (abortion probability) $\Pr[\mathscr{P}_D(X,Y) = \perp] \leq \beta$, and
2. (failure probability) $\Pr[\mathscr{P}_D(X,Y) \neq f(X,Y) \mid \mathscr{P}_D(X,Y) \neq \perp] \leq \delta$.

If $\mathscr{P}$ is a randomized protocol for f, we view it as a distribution over deterministic protocols and we say that it (α, β, δ)-computes f with respect to μ if $\Pr_{\mathscr{P}_D \sim \mathscr{P}}[\mathscr{P}_D$ (β, δ)-computes f w.r.t. $\mu] \geq 1 - \alpha$.

We also need to define an appropriate notion of *conditional* information complexity for protocols with abortion, for which we shall use the notation $\mathrm{IC}^{\mu}_{\alpha,\beta,\delta}(f \mid \nu)$. Let f be as above and let λ be a distribution over the augmented space $\mathscr{X} \times \mathscr{Y} \times \mathscr{D}$, where $\mathscr{D}$ is some finite set. Then λ induces marginals μ on $\mathscr{X} \times \mathscr{Y}$ and ν on $\mathscr{D}$. We say that ν *partitions* μ if, with $(X,Y,D) \sim \lambda$ and $d \in \mathrm{supp}(D)$, the distribution of (X,Y) conditioned on $D = d$ is a product distribution.

Definition 3.11 (Conditional information complexity)**.** Let $\mathscr{P}, f, \lambda$ be as above. The *conditional information cost* of $\mathscr{P}$ under λ is defined as $\mathrm{I}(\mathscr{P}(X,Y) : X, Y \mid D)$, where $(X,Y,D) \sim \lambda$. The *conditional information complexity* of f with respect to λ, denoted by $\mathrm{IC}_{\mu,\delta}(f \mid \nu)$, is defined as the infimum of conditional information costs of protocols that compute f with worst-case error at most δ.

The *information complexity of f with aborts*, denoted by $\mathrm{IC}^{\mu}_{\alpha,\beta,\delta}(f \mid \nu)$, is the infimum of conditional information costs of protocols with abortion that (α, β, δ)-compute f. The bounded-round analogs $\mathrm{IC}^{\mu,(r)}_{\delta}(f \mid \nu)$ and $\mathrm{IC}^{\mu,(r)}_{\alpha,\beta,\delta}(f \mid \nu)$ are defined by taking the respective infimums over only r-round protocols.

4 Upper Bounds for Equality

In this section, we provide deterministic and randomized protocols for EQ_n with low refutation cost and low verification cost. Recall Definition 2.4, which introduced the quantity $\hat{n} = \min\left\{n + \log(1-\delta), \log\frac{(1-\delta)^2}{\varepsilon}\right\}$ as the effective instance size. One can derive one-sided-error and zero-error versions of these results by setting δ and/or ε to zero as needed, and using the convention $\log(w/0) = +\infty$ for $w > 0$. One can in fact tighten the analysis for the case $\varepsilon = \delta = 0$ to obtain the bounds in Theorem 2.3.

Theorem 4.1. *Suppose $n, r \in \mathbb{N}$ and $\varepsilon, \delta \in [0,1]$ are such that $\delta < 1 - 2^{-n/2}$ and $\operatorname{ilog}^{r-1}\hat{n} \geq 4$. Then*

$$\mathrm{D}^{\mu_u,(r),\mathrm{ref}}_{\varepsilon,\delta}(\mathrm{EQ}_n) \leq (1-\delta)\operatorname{ilog}^{r-1}\hat{n} + 5\,.$$

Proof. To gain intuition, we first consider $\delta = 0$, in which case we have $\hat{n} = \min\{n, \log(1/\varepsilon)\}$. The basic idea was already outlined in Section 1. Since we need only handle a random input, we do not need fingerprints. Instead, Alice and Bob take turns revealing increasingly longer prefixes of their inputs: in the jth round, the player to speak sends the next $\approx \operatorname{ilog}^{r-j}\hat{n}$ bits of her input. Whenever a player witnesses a mismatch in prefixes, she *cuts off* the protocol (and the protocol outputs 0). If the protocol ends without a cutoff, it outputs 1. The protocol described so far clearly has no false negatives, and after filling in some details (see below), we can show that it has the desired refutation cost and refutation error.

To achieve further savings for nonzero δ, we partition $\{0,1\}^n$ into sets $S, T \subseteq \{0,1\}^n$ such that $|S| \approx (1-\delta)2^n$. Each player cuts off the protocol at her first opportunity if her input lies in T. Otherwise, they emulate the above protocol on the smaller input space $S \times S$.

We now describe our protocol precisely. Set

$$\begin{aligned}
n' &:= n + \lceil \log(1-\delta) \rceil\,, \\
n'' &:= \min\{n', 2 + \lceil \log((1-\delta)^2/\varepsilon) \rceil\}\,, \\
t_j &:= \begin{cases} \lceil \operatorname{ilog}^{r-j}\hat{n} \rceil\,, & \text{if } 1 \leq j < r, \\ n'' - \sum_{j=1}^{r-1} t_j\,, & \text{if } j = r\,. \end{cases}
\end{aligned}$$

Choose an arbitrary partition of $\{0,1\}^n$ into subsets S and T such that $|S| = 2^{n'}$. Fix an arbitrary bijection $g : S \to \{0,1\}^{n'}$.

The protocol—which we call $\mathscr{P}$—works as follows on input $(x,y) \in \{0,1\}^n \times \{0,1\}^n$. We write $x[i_1 : i_2]$ to denote the substring $x_{i_1} x_{i_1+1} \ldots x_{i_2}$ of x. Each nonempty message in the protocol will be either the string "0", indicating cutoff, or "1" followed by a *payload* string. Each player maintains a variable ℓ that records the length of the prefix that has been compared so far; initially they set $\ell \leftarrow 0$.

The players keep track of whether a cutoff has occurred. Once a cutoff occurs, all further messages in the protocol will be empty strings. Once r rounds have been completed, the appropriate player will output 0 if a cutoff has occurred, and 1 otherwise.

Round j proceeds as follows. Let $P \in \{\text{Alice, Bob}\}$ be the player who speaks in this round, and let $z \in \{x,y\}$ be their input. If necessary, P cuts off if $z \in T$. Now suppose that a cutoff has not yet occurred. If $j = 1$, then P sends the substring $g(z)[1 : t_1]$, sets $\ell \leftarrow t_1$, and the round ends. Otherwise, suppose P receives a non-cutoff message with payload w. If P finds that $w \neq g(z)[\ell+1 : \ell+t_{j-1}]$ then she cuts off the protocol, else if $j < r$, she continues the protocol by sending 1 followed by the next t_j bits of $g(z)$, i.e., she sends $g(z)[\ell+t_{j-1}+1 : \ell+t_{j-1}+t_j]$, sets $\ell \leftarrow \ell+t_{j-1}+t_j$, and the round ends.

The protocol's logic is shown in pseudocode form below, for readers who prefer that presentation.

It is easy to see that $\mathrm{verr}^{\mu_u}(\mathscr{P}) \leq \delta$, since players only cut off on an (x,x) input when $x \in T$. Next, note that a false positive occurs only when $(x,y) \in S \times S$ and $g(x)[1 : n''] = g(y)[1 : n'']$. When $n'' = n'$ (which

Algorithm 1: Round j of the protocol $\mathscr{P}$. Here $t_0 = 0$ and "Round $r+1$" is the output announcement.

```
if j ≤ r then
    if cutoff then send emptystring;
    else
        if z ∈ T then cutoff;
        w ← payload of most recently received message;
        if w ≠ g(z)[ℓ+1 : ℓ+t_{j-1}] then cutoff;
        send "1" followed by g(x)[ℓ+t_{j-1}+1 : ℓ+t_{j-1}+t_j], and set ℓ ← ℓ+t_{j-1}+t_j ;
else
    if cutoff then output 0;
    else
        w ← payload of most recently received message;
        if w ≠ g(z)[ℓ+1 : ℓ+t_{j-1}] then output 0;
        else output 1;
```

corresponds, roughly, to $\varepsilon < (1-\delta)2^{-n}$), Alice and Bob end up comparing all bits of $g(x)$ and $g(y)$, and we get $\mathrm{rerr}^{\mu_u}(\mathscr{P}) = 0$. In the other case, we have $n'' = 2 + \lceil \log((1-\delta)^2/\varepsilon) \rceil$. Letting $(X,Y) \sim \mu_u$, we have

$$\begin{aligned}\mathrm{rerr}^{\mu_u}(\mathscr{P}) &= \Pr[(X,Y) \in S \times S \mid X \neq Y] \cdot \Pr\left[g(X)[1:n''] = g(Y)[1:n''] \mid g(X) \neq g(Y)\right] \\ &\leq \left(2^{n'-n}\right)^2 \cdot \frac{2^{n'-n''}-1}{2^{n'}-1} \leq 2^{2\lceil \log(1-\delta) \rceil} \cdot 2^{-n''} \leq 2^{2(1+\log(1-\delta))} \cdot \frac{\varepsilon}{4(1-\delta)^2} = \varepsilon\,.\end{aligned}$$

Finally, we analyze the refutation cost. Let a_j denote the expected total communication in rounds $\geq j$, conditioned on not cutting off before round j. For convenience, set $a_{r+1} = 0$. We claim that $a_j \leq 3$ for all $j > 2$ and prove so by induction from $r+1 \rightsquigarrow 3$. The base case ($j = r+1$) is trivial. Conditioned on not cutting off before the jth round, the player whose turn it is to speak receives t_{j-1} bits to compare with her own input. Estimating as above, this will fail to cause a cutoff with probability at most $2^{-t_{j-1}}$. Therefore, the player to speak will send at most 1 bit in this round to indicate cutoff (or not) plus, with probability at most $2^{-t_{j-1}}$, will continue the communication, which will cost t_j bits in this round and a_{j+1} bits in expectation in subsequent rounds. The net result is that

$$a_j \leq 1 + 2^{-t_{j-1}}(t_j + a_{j+1}) \leq 1 + \frac{1}{\operatorname{ilog}^{r-j} \hat{n}} \left(\lceil \operatorname{ilog}^{r-j} \hat{n} \rceil + 3 \right) \leq 2 + \frac{4}{\operatorname{ilog}^{r-j} \hat{n}} \leq 3\,.$$

The first two rounds are slightly different, because each player summarily cuts off when her input lies in T. In the first round, Alice cuts off with probability at most δ. In the second round, conditioned on Alice not cutting off, Bob cuts off with probability all but $(1-\delta)2^{-t_1}$. The refutation cost of r-round protocols is therefore bounded by

$$\begin{aligned}\mathrm{rcost}^{\mu_u}(\mathscr{P}) = a_1 &\leq 1 + (1-\delta)t_1 + (1-\delta)\left(1 + (1-\delta)2^{-t_1}(t_2 + a_3)\right) \\ &\leq 1 + (1-\delta)(\lceil \operatorname{ilog}^{r-1} \hat{n} \rceil + 1) + (1-\delta)^2 \frac{\lceil \operatorname{ilog}^{r-2} \hat{n} \rceil + 3}{\operatorname{ilog}^{r-2} \hat{n}} \\ &\leq 1 + (1-\delta)\operatorname{ilog}^{r-1} \hat{n} + 2(1-\delta) + (1-\delta)^2 \left(1 + \frac{4}{\operatorname{ilog}^{r-2} \hat{n}}\right) \\ &\leq 1 + (1-\delta)\operatorname{ilog}^{r-1} \hat{n} + 2(1-\delta) + 2(1-\delta)^2 \\ &\leq 5 + (1-\delta)\operatorname{ilog}^{r-1} \hat{n}\,.\end{aligned}$$

□

Theorem 4.2. *With $n, r, \varepsilon, \delta$ as above, we have $\mathrm{D}^{\mu_u,(r),\mathrm{ver}}_{\varepsilon,\delta}(\mathrm{EQ}_n) \leq (1-\delta)\hat{n} + 3$.*

Proof. We construct a *one-round* protocol achieving the stated verification cost, using S, T, g as in Theorem 4.1. On input (x,y), Alice cuts off if $x \in T$. Otherwise, she sends Bob a prefix of $g(x)$ of length $\min\{n + \lceil \log(1-\delta) \rceil, 2 + \lceil \log((1-\delta)^2/\varepsilon) \rceil$. Bob outputs 0 ("unequal") if (i) Alice cut off, (ii) $y \in T$, or (iii) Alice's prefix does not match that of $g(y)$.

As in the previous proof, this protocol—call it $\mathscr{Q}$—only produces false negatives when inputs lie in T, so that $\mathrm{verr}^{\mu_u}(\mathscr{Q}) \le \delta$. And as before, we get $\mathrm{rerr}^{\mu_u}(\mathscr{Q}) = 0$ for small ε and $\mathrm{rerr}^{\mu_u}(\mathscr{Q}) \le 2^{2\lceil \log(1-\delta) \rceil} \cdot \frac{\varepsilon}{4(1-\delta)^2} \le \varepsilon$ otherwise. As for verification cost, the protocol always sends a bit to indicate cutoff (or not), and for all $(x,x) \in S \times S$ the protocol sends at most $\hat{n}+2$ bits. Thus, $\mathrm{vcost}^{\mu_u}(\mathscr{Q}) \le 1 + (1-\delta)(\hat{n}+2) \le (1-\delta)\hat{n} + 3$. □

Theorem 4.3. *Let $\mathscr{P}$ be an r-round deterministic protocol for EQ_n. Then, there exists an r-round randomized protocol $\mathscr{Q}$ for EQ_n with $\mathrm{verr}(\mathscr{Q}) = \mathrm{verr}^{\mu_u}(\mathscr{P})$, $\mathrm{rerr}(\mathscr{Q}) = \mathrm{rerr}^{\mu_u}(\mathscr{P})$, $\mathrm{rcost}(\mathscr{Q}) = \mathrm{rcost}^{\mu_u}(\mathscr{P})$, and $\mathrm{vcost}(\mathscr{Q}) = \mathrm{vcost}^{\mu_u}(\mathscr{P})$.*

Proof. Construct $\mathscr{Q}$ as follows. Alice and Bob use public randomness to generate a uniform bijection $G : \{0,1\}^n \to \{0,1\}^n$. On input (x,y), they run $\mathscr{P}$ on $(G(x), G(y))$. Note that if $x = y$ then $(G(x), G(y))$ is uniform over $\mathrm{EQ}_n^{-1}(1)$, and if $x \ne y$ then $(G(x), G(y))$ is uniform over $\mathrm{EQ}_n^{-1}(0)$. Thus, distributional guarantees for $\mathscr{P}$ under the uniform distribution become worst-case guarantees for $\mathscr{Q}$. □

Together with Theorems 4.1 and 4.2, this gives upper bounds for randomized protocols.

Corollary 4.4. $\mathrm{R}_{\varepsilon,\delta}^{(r),\mathrm{ref}}(\mathrm{EQ}_n) \le (1-\delta)\,\mathrm{ilog}^{r-1}\hat{n} + 5$.

Corollary 4.5. $\mathrm{R}_{\varepsilon,\delta}^{(r),\mathrm{ver}}(\mathrm{EQ}_n) \le (1-\delta)\hat{n} + 3$.

5 Bounded-Round Communication Lower Bounds for Equality

In this section, we prove all of our communication cost lower bounds on EQ_n. We deal with information cost in the next section. We think of these lower bounds as "combinatorial" (as opposed to "information theoretic"). An important ingredient in some of these combinatorial lower bounds is the *round elimination* technique, which dates back to the work of Miltersen et al. [41].

The proofs in this section will use Kraft's Inequality (Fact 3.5), Shannon's Source Coding Theorem (Fact 3.6), as well as the following approximation lemma.

Lemma 5.1. *For $a \le 2^{n/2}$, $t \le \log^* n - 2$, and $x \in \left[\frac{1}{a}, 1\right]$, we have $\mathrm{ilog}^{t-1} n \ge \mathrm{ilog}^t(2^n x) \ge \left(1 - \frac{\log a}{n}\right)\mathrm{ilog}^{t-1} n$.*

Proof. The upper bound is trivial. We prove the lower bound by induction on t. We have $\log(2^n x) = n + \log x \ge n - \log a > \left(1 - \frac{\log a}{n}\right)n$, and the claim holds for $t = 1$. For $t > 1$, we have

$$\begin{aligned}
\mathrm{ilog}^t(2^n x) &\ge \log\left(1 - \frac{\log a}{n}\right) + \log\left(\mathrm{ilog}^{t-2} n\right) && \text{[by induction hypothesis]} \\
&\ge -\frac{2\log a}{n} + \mathrm{ilog}^{t-1} n && \text{[using } 1 - w \ge 2^{-2w} \text{ for } 0 \le w \le 1/2\text{]} \\
&\ge \left(1 - \frac{\log a}{n}\right)\mathrm{ilog}^{t-1} n && \text{[using } \mathrm{ilog}^{t-1} n \ge 2\text{]}\,.
\end{aligned}$$

□

5.1 Lower Bounds for Zero-Error Protocols

In this section, we provide nearly exact bounds for zero-error protocols.

Theorem 5.2. *For all $r < \log^* n$ we have $\mathrm{D}_{0,0}^{\mu_u,(r),\mathrm{ref}}(\mathrm{EQ}_n) \ge \mathrm{ilog}^{r-1} n - 1$.*

To prove this theorem, we must analyze EQUALITY protocols on finite sets of arbitrary size. Given a finite set S, define EQ_S to be the EQUALITY problem, but when $x, y \in S$. In the following theorem, we let μ_u be uniform over $S \times S$.

Theorem 5.3. *For all integers $r > 0$, we have $\mathrm{D}^{\mu_u,(r),\mathrm{ref}}_{0,0}(\mathrm{EQ}_S) \geq \operatorname{ilog}^r |S| - 1$.*

Proof. Assume $\operatorname{ilog}^r |S| > 1$ as otherwise there is nothing to prove. Define $m = \log |S|$. It might be helpful to think of m as an integer, but this is not necessary.

The proof proceeds by induction on r. When $r = 1$, Alice must send her entire input to achieve zero error in a single round. This costs $\lceil m \rceil > \operatorname{ilog}^1 |S| - 1$ bits, and the theorem holds. Now, assume $\mathrm{D}^{\mu_u,(\ell),\mathrm{ref}}_{0,0}(\mathrm{EQ}_T) \geq \operatorname{ilog}^\ell |T| - 1$ for all finite sets T, and let $\mathscr{P}$ be an optimal $(\ell+1)$-round deterministic protocol for EQ_S. We aim to show that $\mathrm{rcost}^{\mu_u}(\mathscr{P}) \geq \operatorname{ilog}^{\ell+1} |S| - 1 = \operatorname{ilog}^\ell m - 1$. Let $\mathfrak{m}_1, \ldots, \mathfrak{m}_t$ be the possible messages Alice sends in the first round of $\mathscr{P}$. For $1 \leq i \leq t$, Let A_i denote the set of inputs on which Alice sends $\mathfrak{m}_i$, and let ℓ_i denote the length of $\mathfrak{m}_i$. Assume without loss of generality that $\ell_1 \leq \ell_2 \leq \cdots \leq \ell_t$. Since $\mathscr{P}$ is optimal, we must have $|A_1| \geq |A_2| \geq \cdots \geq |A_t|$: otherwise, we can permute which messages are sent on which sets A_i and reduce the overall cost of the protocol.

We analyze the cost of $\mathscr{P}$ by conditioning on Alice's first message. Under the uniform distribution, Alice sends $\mathfrak{m}_i$ with probability $p_i := |A_i|/2^m$. If $y \notin A_i$, Bob refutes equality and the protocol aborts. Thus, over $x \neq y$ inputs, the probability that Bob aborts is $(|A_i| - 1)/(2^m - 1)$. Furthermore, conditioned on the events that (i) Alice's first message is $\mathfrak{m}_i$ and that (ii) Bob doesn't abort, Alice and Bob's inputs are each uniform over A_i. Thus, the remaining communication is at least $\mathrm{D}^{\mu_u,(\ell),\mathrm{ref}}_{0,0}(\mathrm{EQ}_{A_i})$.

Fix $\tau := 2/\operatorname{ilog}^{\ell-1} m$. Call the ith message *small* if $p_i \leq \tau$ and *large* otherwise. We bound

$$\begin{aligned}
\mathrm{rcost}^{\mu_u}(\mathscr{P}) &= \sum_{1 \leq i \leq t} p_i \Big(\ell_i + \frac{|A_i| - 1}{2^m - 1} \mathrm{D}^{\mu_u,(\ell),\mathrm{ref}}_{0,0}(\mathrm{EQ}_{A_i}) \Big) \\
&\geq \sum_{1 \leq i \leq t} p_i \Big(-\log p_i + (p_i - 2^{-m}) \mathrm{D}^{\mu_u,(\ell),\mathrm{ref}}_{0,0}(\mathrm{EQ}_{A_i}) \Big) \\
&\geq \sum_{\text{small } \mathfrak{m}_i} p_i(-\log p_i) + \sum_{\text{large } \mathfrak{m}_i} p_i \Big(-\log p_i + (p_i - 2^{-m})(\operatorname{ilog}^\ell |A_i| - 1) \Big) \\
&\geq \Pr[\text{small message}] \cdot (\operatorname{ilog}^\ell(m) - 1) + \sum_{\text{large } \mathfrak{m}_i} p_i \Big(-\log p_i + p_i \operatorname{ilog}^\ell |A_i| - p_i - 1 \Big) \\
&= \Pr[\text{small message}] \cdot (\operatorname{ilog}^\ell(m) - 1) + \sum_{\text{large } \mathfrak{m}_i} p_i f(p_i) \,,
\end{aligned}$$

where we define $f(x) := -\log x + x \operatorname{ilog}^\ell(2^m x) - x - 1$. The first inequality holds by the source coding theorem (Fact 3.6) and the third inequality holds because $p_i \leq \tau$ for all small messages.

We now claim that $f'(x) > 0$ for all $x \in [\tau, 1]$. We prove this by explicitly calculating the derivative of f. If $x \geq \tau$, then $-1/(x \ln 2) \geq -\operatorname{ilog}^{\ell-1}(m)/(2 \ln 2)$. By Lemma 5.1, we have

$$\begin{aligned}
f'(x) &= -\frac{1}{x \ln 2} + \operatorname{ilog}^\ell(2^m x) - \frac{1}{(\ln 2)(\ln x \cdot 2^m) \prod_{j=0}^{\ell-2} \ln(\operatorname{ilog}^j x \cdot 2^m)} - 1 \\
&\geq -\frac{\operatorname{ilog}^{\ell-1} m}{2 \ln 2} + \operatorname{ilog}^{\ell-1} m - \frac{(\operatorname{ilog}^{\ell-1} m) \operatorname{ilog}^\ell m}{m} - o(1) - 1 \\
&= \big(\operatorname{ilog}^{\ell-1} m\big) \left(1 - \frac{1}{2 \ln 2}\right) - 1 - o(1) = \Omega(\operatorname{ilog}^{\ell-1} m) \,,
\end{aligned}$$

which proves the claim. It now follows that for large messages, $f(p_i)$ is minimized at $f(\tau)$. Note that

$$\begin{aligned} f(\tau) &= -\log\tau + \tau\,\mathrm{ilog}^{\ell}(2^m\tau) - \tau - 1 \\ &\geq \mathrm{ilog}^{\ell} m - 1 + \frac{2}{\mathrm{ilog}^{\ell-1} m}\,\mathrm{ilog}^{\ell-1} m\Big(1 - \frac{\mathrm{ilog}^{\ell}(m)-1}{m}\Big) - \frac{2}{\mathrm{ilog}^{\ell-1} m} - 1 \\ &> \mathrm{ilog}^{\ell} m - 1\,. \end{aligned}$$

Plugging this back into our inequality for the cost of $\mathscr{P}$, we get

$$\mathrm{rcost}^{\mu_u}(\mathscr{P}) \geq \Pr[\text{small message}]\cdot(\mathrm{ilog}^{\ell} m - 1) + \Pr[\text{large message}]\cdot(\mathrm{ilog}^{\ell} m - 1) = \mathrm{ilog}^{\ell} m - 1\,. \qquad \square$$

Theorem 5.4. $\mathrm{D}_{0,0}^{\mu_u,(r),\mathrm{ver}}(\mathrm{EQ}_n) \geq n$. *Note that this lower bound is independent of r.*

Proof. Let $\mathscr{P}$ be a deterministic zero-error protocol for EQ_n. As the protocol has no error, the communication matrix is partitioned into monochromatic rectangles. In particular, there are 2^n 1-rectangles, since each (x,x) input must map to a different rectangle.[6] Let T_x and ℓ_x denote the protocol transcript corresponding to (x,x) and the length of this protocol transcript, respectively. Note that $\{T_x\}$ form a prefix-free coding of $\{0,1\}^n$. By Kraft's inequality, we have $\sum_x 2^{-\ell_x} \leq 1$. Therefore, in expectation $\mathbb{E}[2^{-\ell_x}] \leq 2^{-n}$, and by Jensen's inequality, we get the following.

$$-n \geq \log\mathbb{E}[2^{-\ell_x}] \geq \mathbb{E}[\log(2^{-\ell_x})] = -\mathbb{E}[\ell_x]\,.$$

Multiplying each side of the inequality by -1, we have $\mathbb{E}_x[\ell_x] \geq n$. This is precisely $\mathrm{vcost}^{\mu_u}(\mathscr{P})$, thus the proof is complete. $\square$

Theorem 5.5. $\mathrm{R}_{0,0}^{(r),\mathrm{ver}}(\mathrm{EQ}_n) \geq n$. *As above, this lower bound is independent of r.*

Proof. Let $\mathscr{P}$ be a randomized zero-error protocol for EQ_n. Given any string s, let $\mathscr{P}_s$ denote the deterministic protocol obtained by fixing the public randomness to s. Proceeding along the same lines as in the proof of Theorem 5.4, we have $\mathbb{E}[\ell_{x,s}] \geq n$, where $\ell_{x,s}$ is the length of the protocol transcript in $\mathscr{P}_s$ on input (x,x). This holds for every $\mathscr{P}_s$, hence $\mathbb{E}_{x,s}[\ell_{x,s}] \geq n$. Therefore, there exists x such that $\mathbb{E}_s[\ell_{x,s}] \geq n$. Recalling the definition of vcost, we have $\mathrm{vcost}(\mathscr{P}) \geq \mathrm{cost}(\mathscr{P};x,x) = \mathbb{E}_s[\ell_{x,s}] \geq n$, completing the proof. $\square$

5.2 Refutation Lower Bounds for Protocols with Two-Sided Error

In this section, we give combinatorial lower bounds on the refutation cost of EQUALITY protocols that admit error. All of the bounds in this section will be asymptotic rather than nearly exact. For this reason, we will strive for simplicity of the proofs at the possible expense of some technical accuracy. For instance, we will often drop ceilings or floors in the mathematical notation. We will also assume that players have the ability to instantly abort a protocol when equality has been refuted. This is easily implemented, as seen in Section 5.1 at negligible communication cost. We prefer to avoid the technical machinery needed to express this explicitly.

Definition 5.6. An $\langle n,r,\varepsilon,\delta,c\rangle$-EQUALITY protocol $\mathscr{P}$ is a r-round deterministic protocol with $\mathrm{rerr}^{\mu_u}(\mathscr{P}) \leq \varepsilon$, $\mathrm{verr}^{\mu_u}(\mathscr{P}) \leq \delta$, and $\mathrm{rcost}^{\mu_u}(\mathscr{P}) \leq c$.

For the sake of brevity, we often drop the "EQUALITY" and simply refer to an $\langle n,r,\varepsilon,\delta,c\rangle$-protocol. Our first lemma demonstrates that disallowing false negatives changes the communication complexity very little.

[6] If (x,x) and (y,y) were in the same rectangle, then so would (x,y) and (y,x). Thus, the protocol would err on these inputs.

Lemma 5.7. *If there exists a $\langle n, r, \varepsilon, \delta, c\rangle$-EQUALITY protocol, then there exists a $\langle n', r, \varepsilon', 0, c'\rangle$-EQUALITY protocol, where $n' = n + \log(1-\delta)$, $\varepsilon' = 2\varepsilon/(1-\delta)^2$, and $c' = 2c/(1-\delta)^2$.*

Proof. Let $S = \{x : \mathrm{out}(\mathscr{P}(x,x)) = 0\}$ be the set of inputs on which $\mathscr{P}$ gives a false negative, and let $T = \{0,1\}^n \setminus S$. Since $\mathscr{P}$ has false negative rate δ under the uniform distribution, we have $|T| \geq (1-\delta)2^n = 2^{n'}$.

First create a new EQ_n protocol $\mathscr{P}'$ which works as follows. On input (x,y), Alice aborts and outputs 0 if $x \in S$; otherwise, the players emulate $\mathscr{P}$ and output $\mathrm{out}(\mathscr{P}(x,y))$. Note that $\mathscr{P}'$ makes precisely the same false negatives as in $\mathscr{P}$, and aborting when $x \in S$ can only decrease the false positive rate and the expected communication on inputs in $\mathrm{EQ}_n^{-1}(0)$. Thus, $\mathscr{P}'$ is also a $\langle n, r, \varepsilon, \delta, c\rangle$-protocol.

Next, fix an arbitrary bijection $g : \{0,1\}^{n'} \to T$, and construct an $\mathrm{EQ}_{n'}$ protocol $\mathscr{Q}$ in the following way. On input (X,Y), players emulate $\mathscr{P}'$ on input $(g(X), g(Y))$ and output $\mathrm{out}(\mathscr{P}'(g(X), g(Y)))$. Note that $g(X), g(Y) \in T$, so there are no false negatives. There can be as many false positives as in $\mathscr{P}'$. However, the sample space is smaller ($2^{2n'} - 2^{n'}$ vs $2^{2n} - 2^n$), so the false positive rate can increase. By Lemma 3.3, the overall error is at most $2\varepsilon/(1-\delta)^2$. Similarly, the communication in $\mathscr{Q}$ on any input (X,Y) is the same as the communication in $\mathscr{P}'$ on input $(g(X), g(Y))$, but since the sample space is smaller (again $2^{2n'} - 2^{n'}$ vs. $2^{2n} - 2^n$), the expected communication can increase. However, the overall increase in communication is at most a factor of $2/(1-\delta)^2$ by Lemma 3.3. □

Lemma 5.8 (Combinatorial Round Elimination for EQUALITY)**.** *If there is an $\langle n, r, \varepsilon, 0, c\rangle$-EQUALITY protocol, then there is an $\langle n - 3c - 2, r-1, 12\varepsilon 2^{3c}, 0, 12c2^{3c}\rangle$-EQUALITY protocol.*

Proof. Let $\mathscr{P}$ be a $\langle n, r, \varepsilon, 0, c\rangle$-protocol. Let $Z(x,y) = 1$ if the protocol errs on input (x,y), and let $Z(x,y) = 0$ otherwise. Then we have

$$\mathbb{E}_x\left[\mathbb{E}_{y\neq x}[|\mathscr{P}(x,y)|]\right] \leq c, \quad \text{and} \quad \mathbb{E}_x\left[\mathbb{E}_{y\neq x}[Z(x,y)]\right] \leq \varepsilon\,.$$

Call x good if (1) $\mathbb{E}_{y\neq x}[|\mathscr{P}(x,y)|] \leq 3c$, and (2) $\mathbb{E}_{y\neq x}[Z(x,y)] \leq 3\varepsilon$. By two applications of Markov's inequality and a union bound, at least $2^n/3$ of the x are good. Next, fix Alice's first message m so it is constant over the maximal number of good x. Any message m sent on a good x must have $|m| < 3c$ (otherwise it would violate the goodness of x.) It follows that m is constant over a set A of good x of size $|A| \geq 2^{n-3c-2}$. This induces a $(r-1)$-round protocol $\mathscr{Q}$ for EQ_A. It remains to bound the cost and error of $\mathscr{Q}$. Applying Lemma 3.3 twice, we have that the cost and error are bounded by (respectively)

$$\mathrm{rcost}^{\mu_u}(\mathscr{Q}) = \mathbb{E}_{x\in A}\left[\mathbb{E}_{y\in A, y\neq x}[|\mathscr{P}(x,y)|]\right] \leq \frac{2^n}{2^{n-3c-2}}\,\mathbb{E}_{x\in A}\left[\mathbb{E}_{y\in\{0,1\}^n, y\neq x}[|\mathscr{P}(x,y)|]\right] \leq 12c2^{3c}\,,$$

$$\mathrm{verr}^{\mu_u}(\mathscr{Q}) = \mathbb{E}_{x\in A}\left[\mathbb{E}_{y\in A, y\neq x}[Z(x,y)]\right] \leq \frac{2^n}{2^{n-3c-2}}\,\mathbb{E}_{x\in A}\left[\mathbb{E}_{y\in\{0,1\}^n, y\neq x}[Z(x,y)]\right] \leq 12\varepsilon 2^{3c}\,.$$ □

Corollary 5.9. *Let n, j, r, d be integers with $n > d$, d sufficiently large, and $r \geq 1$. Suppose there exists an $\langle n, r, \varepsilon\ell, 0, \ell\rangle$-protocol, where $\ell = \frac{1}{6}\,\mathrm{ilog}^j d$. Then, there exists an $\langle n - 3\ell - 2, r-1, \varepsilon\ell', 0, \ell'\rangle$-protocol with $\ell' = \frac{1}{6}\,\mathrm{ilog}^{j-1} d$.*

Proof. This boils down to the following estimations, which are valid for all sufficiently large d.

$$12\ell 2^{3\ell} = 2(\mathrm{ilog}^j d)2^{\frac{1}{2}\,\mathrm{ilog}^j d} = 2\,\mathrm{ilog}^j d\sqrt{\mathrm{ilog}^{j-1} d} < \frac{1}{6}\,\mathrm{ilog}^{j-1} d\,.$$ □

Theorem 5.10 (Lower Bound for Protocols with False Negatives Disallowed)**.** *Let n be a sufficiently large integer, $\varepsilon < 1/4$ a real, and $r \geq 1$. Fix $\tilde{n} := \min\{n, \log(1/\varepsilon)\}$. Then, $\mathrm{D}^{\mu_u,(r),\mathrm{ref}}_{\varepsilon,0}(\mathrm{EQ}_n) = \Omega(\mathrm{ilog}^{r-1}\tilde{n})$.*

Proof. In this proof we tacitly assume $\mathrm{ilog}^{r-1}\tilde{n} \geq 100$.

Suppose for the sake of a contradiction that there exists a $\langle n, r, \varepsilon, 0, \frac{1}{6}\operatorname{ilog}^{r-1}\tilde{n}\rangle$-protocol $\mathscr{P}$. Applying Lemma 5.8 gives an $\langle n - \frac{3}{5}\operatorname{ilog}^{r-1}\tilde{n}, r-1, \frac{\varepsilon}{6}\operatorname{ilog}^{r-2}\tilde{n}, 0, \frac{1}{6}\operatorname{ilog}^{r-2}\tilde{n}\rangle$-protocol $\mathscr{P}'$. Next, applying Corollary 5.9 repeatedly, a total of $r-2$ times, gives an $\langle n - \frac{3}{5}\sum_{j=1}^{r-1}\operatorname{ilog}^{j}\tilde{n}, 1, \frac{\varepsilon}{6}\tilde{n}, 0, \frac{\tilde{n}}{6}\rangle$-protocol. Finally, applying Lemma 5.8 once more gives an $\langle n - \frac{3}{5}\sum_{j=0}^{r-1}\operatorname{ilog}^{j}\tilde{n}, 0, 2\varepsilon\tilde{n}2^{\tilde{n}/2}, 0, 2\tilde{n}2^{\tilde{n}/2}\rangle$-protocol $\mathscr{Q}$.

Note that since $\mathscr{Q}$ has false negative rate zero, $\mathscr{Q}$ must output 1 with certainty. Thus, $\mathscr{Q}$ errs on all $X \neq Y$ inputs; i.e., $\mathscr{Q}$ has false positive rate 1. On the other hand, $\tilde{n} \leq \log(1/\varepsilon)$, so the false positive rate of $\mathscr{Q}$ is $2\varepsilon\tilde{n}2^{\tilde{n}/6} \leq \sqrt{\varepsilon} < 1/2$. This is a contradiction as long as the problem remains nontrivial.

Since $\operatorname{ilog}^{j}\tilde{n} \geq 100$, we have $\sum_{j=t+1}^{r-1}\operatorname{ilog}^{j}\tilde{n} < \frac{1}{5}\operatorname{ilog}^{t}\tilde{n}$. Also notice that since $\tilde{n} \leq n$, we have $n - \frac{3}{5}\sum_{j=0}^{r-1}\operatorname{ilog}^{j}\tilde{n} > n/5$. Thus, we have a zero-round protocol for $\mathrm{EQ}_{n'}$ for some $n' = \Omega(n)$ that has false positive rate $< 1/2$ but must output 1 with certainty, a contradiction. □

Theorem 5.11 (Lower Bound for Protocols with Two-Sided Error)**.** *Let n be a sufficiently large integer, and let ε, δ be reals such that $\delta \leq 1 - 2^{-n/2}$ and $\varepsilon/(1-\delta)^2 < 1/8$. Let $\hat{n}$ be as given in Definition 2.4. Then, $\mathrm{D}^{\mu_u,(r),\mathrm{ref}}_{\varepsilon,\delta}(\mathrm{EQ}_n) = \Omega((1-\delta)^2\operatorname{ilog}^{r-1}\hat{n})$.*

Proof. Fix $d = \min\{n/2, \log((1-\delta)^2/2\varepsilon)\}$, so that $\log d = \Theta(\log\hat{n})$. Suppose, to the contrary, that there exists an $\langle n, r, \varepsilon, \delta, \frac{1}{12}(1-\delta)^2\operatorname{ilog}^{r-1}d\rangle$-protocol $\mathscr{P}$. Since $n + \log(1-\delta) > n/2$, Lemma 5.7 gives an $\langle n/2, r, 2\varepsilon/(1-\delta)^2, 0, \frac{1}{6}\operatorname{ilog}^{r-1}d\rangle$-protocol. The rest of the proof echoes the proof of Theorem 5.10. □

Next, we prove a combinatorial lower bound for randomized communication complexity.

Theorem 5.12. *Let n be a sufficiently large integer, ε and δ reals such that $\delta < 1 - 2^{1-n/2}$ and $64\varepsilon < (1-\delta)^3$. Then, $\mathrm{R}^{(r),\mathrm{ref}}_{\varepsilon,\delta}(\mathrm{EQ}_n) = \Omega((1-\delta)^3\operatorname{ilog}^{r-1}\hat{n})$, where $\hat{n}$ is as in Definition 2.4.*

Proof. Let $\mathscr{P}$ be an r-round randomized protocol with $\mathrm{rerr}(\mathscr{P}) = \varepsilon, \mathrm{verr}(\mathscr{P}) = \delta$, and $\mathrm{rcost}^{\mu_u}(\mathscr{P}) = c$. Define $z = 1-\delta, \hat{\varepsilon} = 4\varepsilon/(1-\delta)$, and $\hat{c} = 4c/(1-\delta)$. Let $\mathscr{P}_s$ denote the deterministic protocol obtained from $\mathscr{P}$ by setting its random string to s. Call a string s good if (i) $\mathrm{verr}^{\mu_u}(\mathscr{P}_s) \leq 1 - z/2$, (ii) $\mathrm{rerr}^{\mu_u}(\mathscr{P}_s) \leq \hat{\varepsilon}$, and (iii) $\mathrm{rcost}^{\mu_u}(\mathscr{P}_s) \leq \hat{c}$. Applying a Markov argument to each of these three conditions, we see that

$$\Pr[s \text{ is bad}] < \frac{1-z}{1-z/2} + \frac{z}{4} + \frac{z}{4} < 1\,,$$

where we used $(1-z)/(1-z/2) < 1 - z/2$. Thus there exists a good string s. Note that $\mathscr{P}_s$ is a $[n, r, \hat{\varepsilon}, \hat{\delta}, \hat{c}]$-protocol, and by Theorem 5.11, $\hat{c} = \Omega((1-\delta)^2\operatorname{ilog}^{r-1}\hat{n})$. Therefore, $c = \Omega((1-\delta)^3\operatorname{ilog}^{r-1}\hat{n})$. □

5.3 Verification Lower Bounds for Protocols with Two-Sided Error

Theorem 5.13. $\mathrm{D}^{\mu_u,(r),\mathrm{ver}}_{\varepsilon,\delta}(\mathrm{EQ}_n) \geq (1-\delta)(\hat{n}-1)$, *where $\hat{n}$ is as in Definition 2.4.*

Proof. Fix a deterministic protocol $\mathscr{P}$ achieving $\mathrm{rerr}^{\mu_u}(\mathscr{P}) = \varepsilon$ and $\mathrm{verr}^{\mu_u}(\mathscr{P}) = \delta$. This protocol naturally partitions the communication matrix for EQ_n into combinatorial rectangles. Let $R_1, \ldots, R_c$ be the rectangles on which $\mathscr{P}$ outputs 1. Let s_i denote the number of (x,x) inputs in R_i. Since $\mathscr{P}$ has false negative rate δ, we have $\sum_i s_i = 2^n(1-\delta)$. Let $p_i = s_i/2^n$ and $q_i = p_i/(1-\delta)$. Notice that p_i is the probability that $(x,x) \in R_i$ for a uniformly chosen x. Similarly, q_i is the probability that $(x,x) \in R_i$ conditioned on $\mathscr{P}$ verifying equality on (x,x). We now analyze the false positive rate. Recall that there are $2^{2n} - 2^n$ total $x \neq y$ inputs. It is easy to see that R_i contains at least $s_i^2 - s_i$ false positives. Therefore, we have

$$\varepsilon \geq \frac{1}{2^{2n}-2^n}\sum_{i=1}^{c}(s_i^2 - s_i) = \sum_{i=1}^{c}\frac{s_i(s_i-1)}{2^n(2^n-1)} \geq \sum_{i=1}^{c}p_i(p_i - 2^{-n}) = -2^{-n}(1-\delta) + \sum_{i=1}^{c}p_i^2\,.$$

Rearranging terms and noting that $q_i = p_i/(1-\delta)$, we have

$$\mathbb{E}[q_i] = \sum_{i=1}^{c} q_i^2 = \frac{1}{(1-\delta)^2}\sum_{i=1}^{c} p_i^2 \le \frac{1}{(1-\delta)^2}\left(\varepsilon + 2^{-n}(1-\delta)\right) = \frac{\varepsilon}{(1-\delta)^2} + \frac{2^{-n}}{(1-\delta)} \le 2\cdot 2^{-\hat{n}}.$$

Next, we analyze the verification cost of $\mathscr{P}$. Let ℓ_i denote the length of the protocol transcript for inputs in the rectangle R_i. Observe that the transcripts $\mathscr{P}(x,x)$ with $\mathrm{out}(\mathscr{P}(x,x)) = 1$ give a prefix-free encoding of the set of rectangles $\{R_1,\dots,R_c\}$. Therefore,

$$\begin{aligned}\mathrm{vcost}^{\mu_u}(\mathscr{P}) &= \sum_{x\in\{0,1\}^n} \frac{|\mathscr{P}(x,x)|}{2^n} \ge \sum_{i=1}^{c} p_i\ell_i = (1-\delta)\sum_{i=1}^{c} q_i\ell_i \ge (1-\delta)\sum_{i=1}^{c} q_i(-\log q_i)\\ &= -(1-\delta)\,\mathbb{E}[\log q_i] \ge -(1-\delta)\log\mathbb{E}[q_i] \ge -(1-\delta)(-\hat{n}+1) = (1-\delta)(\hat{n}-1)\,,\end{aligned}$$

where the second inequality is from the source coding theorem (Fact 3.6) and the third is from Jensen's inequality. □

Theorem 5.14. $\mathrm{R}^{(r),\mathrm{ver}}_{\varepsilon,\delta}(\mathrm{EQ}_n) > \frac{1}{8}(1-\delta)^2(\hat{n}+\log(1-\delta)-5)$.

Proof. Suppose there exists a randomized protocol $\mathscr{P}$ with $\mathrm{rerr}(\mathscr{P}) \le \varepsilon$, $\mathrm{verr}(\mathscr{P}) \le \delta$, and $\mathrm{vcost}(\mathscr{P}) \le m$. For a string s, let $\mathscr{P}_s$ denote the deterministic protocol obtained from $\mathscr{P}$ by fixing the public randomness to s. By the cost and error guarantees of $\mathscr{P}$, for all $(x,y) \in \mathrm{EQ}_n^{-1}(1)$ we have $\mathbb{E}_s[\mathrm{cost}(\mathscr{P}_s;x,y)] \le m$ and $\mathbb{E}_s[\Pr[\mathrm{out}(\mathscr{P}_s(x,y)) = 0]] \le \delta$, while for $(x,y) \in \mathrm{EQ}^{-1}(0)$ we have $\mathbb{E}_s[\Pr[\mathrm{out}(\mathscr{P}_s(x,y)) = 1]] \le \varepsilon$. In particular, letting $(X,Y) \sim \mu_u$, we have

$$\begin{aligned}\mathbb{E}_{s,X,Y}[\Pr[\mathrm{out}(\mathscr{P}_s(X,Y)) = 1 \mid X \ne Y]] &\le \varepsilon\,,\\ \mathbb{E}_{s,X,Y}[\Pr[\mathrm{out}(\mathscr{P}_s(X,Y) = 0 \mid X = Y]] &\le \delta\,,\\ \mathbb{E}_{s,X,Y}[\mathrm{cost}(\mathscr{P}_s;X,Y) \mid X = Y] &\le m\,.\end{aligned}$$

Define $z = 1-\delta$, $\hat{\varepsilon} = 4\varepsilon/(1-\delta)$, $\hat{\delta} = 1 - z/2$, and $\hat{m} = 4m/(1-\delta)$. Call a string s good if (i) $\mathrm{verr}(\mathscr{P}_s) \le 1 - z/2$, (ii) $\mathrm{rerr}(\mathscr{P}_s) \le \hat{\varepsilon}$, and (iii) $\mathrm{vcost}^{\mu_u}(\mathscr{P}) \le \hat{m}$. Applying a Markov argument to each condition,

$$\Pr[s \text{ is bad}] < \frac{1-z}{1-z/2} + \frac{z}{4} + \frac{z}{4} < 1\,,$$

where we used $(1-z)/(1-z/2) < 1 - z/2$. Thus, there exists a good string s. Note that $\mathscr{P}_s$ is a deterministic $(\hat{\varepsilon},\hat{\delta})$-error EQ_n protocol. Using Definition 2.4 to figure the new effective instance size and applying Theorem 5.13, we obtain

$$\frac{4m}{1-\delta} \ge \mathrm{vcost}^{\mu_u}(\mathscr{P}_s) \ge \frac{z}{2}\left(\min\left\{n + \log(z/2), \log\frac{z(z/2)^2}{4\varepsilon}\right\} - 1\right) \ge \frac{z}{2}(\hat{n} + \log z - 5)\,.$$

The proof is completed by rearranging the above inequality and substituting $z = 1-\delta$. □

The analysis in the above proof is very loose when δ is bounded away from 1. In particular, when there are no false negatives (i.e., when $\delta = 0$), we are able to show that $\mathrm{R}^{(r),\mathrm{ver}}_{\varepsilon,0} \ge c\hat{n}$ for every constant $c < 1$.

6 Bounded-Round Information Complexity of Equality

In this section we prove Theorem 2.6, which we think of as the most important result of this paper. We wish to lower bound the bounded-round information complexity of EQUALITY with respect to the uniform distribution. Recall that we are concerned chiefly with protocols that achieve very low refutation error,

though they may have rather high verification error. We will prove our lower bound by proving a round elimination lemma for EQ_n that targets *information* cost, and then applying this lemma repeatedly.

This proof has much more technical complexity than our earlier lower bound proofs. Let us see why. There are two main technical difficulties and they arise, ultimately, from the same source: the inability to use (the easy direction of) Yao's minimax lemma. When proving a lower bound on *communication* cost, Yao's lemma allows us to fix the random string used by any purported protocol, which immediately moves us into the clean world of deterministic protocols. This hammer is unavailable to us when working with *information* cost. The most we can do is to "average away" the public randomness. We then have to deal with private coin randomized protocols the entire way through the round elimination argument. As a result, our intermediate protocols, obtained by eliminating some rounds of our original protocol, do not obey straightforward cost and error guarantees. This is the first technical difficulty, and our solution to it leads us to the concept of a "kernel" in Definition 6.1 below.

The second technical difficulty is that we are unable to switch to the simpler case of zero verification error like we did in the proof of Theorem 2.5, Parts (9) and (10). Therefore, all our intermediate protocols continue to have verification error. Since errors scale up with each round elimination, and the verification error starts out high, we cannot afford even a constant-factor scaling. We must play very delicately with our error parameters, which leads us to the somewhat complicated parametrization seen in Definition 6.2 below.

6.1 The Round Elimination Argument

Definition 6.1 (Kernel)**.** Let p and q be probability distributions on $\{0,1\}^n$, let $S \subseteq \{0,1\}^n$, and let $\ell \geq 0$ be a real number. The triple (p,q,S) is defined to be an *ℓ-kernel* if the following properties hold.

[K1] $\text{H}(p) \geq n - \ell$ and $\text{H}(q) \geq n - \ell$.

[K2] $p(S) \geq 2^{-\ell}$ and $q(S) \geq \frac{1}{2}$.

[K3] For all $x \in S$ we have $q(x) \geq 2^{-n-\ell}$.

Some intuition about kernels might be helpful here. Recall that in the combinatorial round elimination lemma of Lemma 5.8, we show that after fixing one round of communication, we are still able to solve EQ on inputs uniformly distributed over a smaller set, albeit with some degredation in cost/error parameters. As mentioned above, we are not able to maintain straightfoward cost and error guarantees in the information-theoretic setting. However, the idea that we can solve EQ when inputs are uniformly distributed over a still-large set should still intuitively hold. Instead of uniformly distributed inputs, we'd like to argue that after eliminating a round of communication, we're able to solve EQ when inputs are *almost* uniform over some smaller set.

Our kernel definition captures enough of this intuition to make the information-theoretic round elimination work. The set S plays the role of the smaller set players will solve EQ on post-round-elimination. Instead of uniform inputs, Alice and Bob's inputs come from some high-entropy *product* distribution. Moreover, the support of these distributions on S is not too low. To maintain the cost/error guarantees, we need Bob's inputs to be reasonably spread out over S. Finally, we need to additionally parameterize *how close to uniform* the input distributions are; this parameter degrades in the round elmination, along with our error guarantees. Nevertheless, we're able to show that as rounds of communication are eliminated, we retain EQ protocols on inputs that remain "reasonably close to uniform" over a reasonably large set S. Our specific protocol parameterization lies below.

Definition 6.2 (Parametrized Protocols)**.** Suppose we have an integer $r \geq 1$, and nonnegative reals ℓ, a, b, and c. A protocol $\mathscr{P}$ for EQ_n is defined to be an $[r,\ell,a,b,c]$-protocol if there exists an ℓ-kernel (p,q,S) such that the following properties hold.

[P1] The protocol $\mathscr{P}$ is private-coin and uses r rounds, with Alice speaking in the first round.

[P2] We have $\mathrm{err}^{p\otimes q|S\times S}(\mathscr{P}) = \Pr_{(X,Y)\sim p\otimes q}[\mathrm{out}(\mathscr{P}(X,Y)) \neq \mathrm{EQ}_n(X,Y) \mid (X,Y)\in S\times S] \leq 2^{-a}$.

[P3] We have $\mathrm{verr}^{p\otimes\xi|S\times S}(\mathscr{P}) = \Pr_{X\sim p}[\mathrm{out}(\mathscr{P}(X,X)) = 0 \mid X\in S] \leq 1-2^{-b}$.

[P4] We have $\mathrm{icost}^{p\otimes q}(\mathscr{P}) \leq c$.

We alert the reader to the fact that [P2] considers overall error, and not refutation error. We encourage the reader to take a careful look at [P3] and verify the equality claimed therein. It is straightforward, once one revisits Definition 2.1 and recalls that ξ denotes the uniform distribution on $\{0,1\}^n$.

Since we have a number of parameters at play, it is worth recording the following simple observation.

Fact 6.3. Suppose that $\ell' \geq \ell, c' \geq c, a' \leq a$, and $b' \geq b$. Then every ℓ-kernel is also an ℓ'-kernel, and every $[r,\ell,a,b,c]$-protocol is also an $[r,\ell',a',b',c']$-protocol.

Theorem 6.4 (Information-Theoretic Round Elimination for EQUALITY)**.** *If there exists an $[r,\ell,a,b,c]$-protocol with $r\geq 1$ and $c\geq 4$, then there exists an $[r-1,\ell',a',b',c']$-protocol, where*

$$\ell' := (c+\ell)2^{\ell+2b+7}, \qquad a' := a-(c+\ell)2^{\ell+2b+8},$$
$$b' := b+2, \qquad c' := (c+2)2^{\ell+2b+6}.$$

Proof. Let $\mathscr{P}$ be an $[r,\ell,a,b,c]$-protocol, and let (p,q,S) be an ℓ-kernel satisfying the conditions in Definition 6.2. Assume WLOG that each message in $\mathscr{P}$ is generated using a fresh random string. Let $X\sim p$ and $Y\sim q$ be independent random variables denoting an input to $\mathscr{P}$. Let $M_1,\ldots,M_r$ be random variables denoting the messages sent in $\mathscr{P}$ on input (X,Y), with M_j being the jth message; note that these variables depend on X,Y, and the random strings used by the players. We then have

$$c \geq \mathrm{icost}^{p\otimes q}(\mathscr{P}) = \mathrm{I}(XY : M_1M_2\ldots M_r) = \mathrm{I}(X:M_1) + \mathrm{I}(XY : M_2\ldots M_r \mid M_1), \tag{4}$$

where the final step uses the chain rule for mutual information, and the fact that M_1 and Y are independent. In particular, we have $\mathrm{I}(X:M_1)\leq c$, and so $\mathrm{H}(X\mid M_1) = \mathrm{H}(X) - \mathrm{I}(X:M_1) \geq n-\ell-c$. By Lemma 3.8,

$$\mathrm{H}(X \mid M_1, X\in S) \geq n - \frac{\ell+c+1}{p(S)} \geq n-(\ell+c+1)2^{\ell}. \tag{5}$$

Let $\mathscr{M}$ be the set of messages that Alice sends with positive probability as her first message in $\mathscr{P}$, given the random input X, i.e., $\mathscr{M} := \{\mathfrak{m} : \Pr[M_1=\mathfrak{m}]>0\}$. Consider a particular message $\mathfrak{m}\in\mathscr{M}$. Let $\mathscr{P}'_{\mathfrak{m}}$ denote the following protocol for EQ_n. The players simulate $\mathscr{P}$ on their input, except that Alice is assumed to have sent $\mathfrak{m}$ as her first message. As a result, $\mathscr{P}'_{\mathfrak{m}}$ has $r-1$ rounds and Bob is the player to send the first message in $\mathscr{P}'_{\mathfrak{m}}$. Let $\pi_{\mathfrak{m}}$ and q' be the distributions of $(X\mid M_1=\mathfrak{m}\wedge X\in S)$ and $(Y\mid Y\in S)$, respectively.

Observe that $\mathrm{icost}^{\pi_{\mathfrak{m}}\otimes q'}(\mathscr{P}'_{\mathfrak{m}}) = \mathrm{I}(XY : M_2\ldots M_r \mid M_1=\mathfrak{m}\wedge(X,Y)\in S\times S)$. Letting L denote a random first message distributed identically to M_1, we now get

$$\begin{aligned}\mathbb{E}_L\left[\mathrm{icost}^{\pi_L\otimes q'}(\mathscr{P}'_L)\right] &= \mathrm{I}(XY : M_2\ldots M_r \mid M_1, (X,Y)\in S\times S)\\ &\leq \frac{\mathrm{I}(XY : M_2\ldots M_r\mid M_1)+1}{p(S)q(S)} \leq (c+1)2^{\ell+1},\end{aligned} \tag{6}$$

where the first inequality uses Lemma 3.7 and the fact that X,Y are independent conditioned on M_1 (since M_1 is a function of X only) and the second inequality uses (4) and Property [K2]. Examining Properties [P2] and [P3], we obtain

$$\mathbb{E}_L\left[\mathrm{err}^{\pi_L\otimes q'}(\mathscr{P}'_L)\right] = \mathrm{err}^{p\otimes q|S\times S}(\mathscr{P}) \leq 2^{-a}, \tag{7}$$
$$\mathbb{E}_L\left[\mathrm{verr}^{\pi_L\otimes\xi}(\mathscr{P}'_L)\right] = \mathrm{verr}^{p\otimes\xi|S\times S}(\mathscr{P}) \leq 1-2^{-b}. \tag{8}$$

Definition 6.5 (Good message)**.** A message $\mathfrak{m} \in \mathscr{M}$ is said to be *good* if the following properties hold:

[G1] $\mathrm{H}(\pi_{\mathfrak{m}}) = \mathrm{H}(X \mid M_1 = \mathfrak{m} \wedge X \in S) \geq n - (\ell + c + 1)2^{\ell+b+3}$,

[G2] $\mathrm{icost}^{\pi_{\mathfrak{m}} \otimes q'}(\mathscr{P}'_{\mathfrak{m}}) \leq 2^{\ell+b+4}(c+1)$,

[G3] $\mathrm{err}^{\pi_{\mathfrak{m}} \otimes q'}(\mathscr{P}'_{\mathfrak{m}}) \leq 2^{-a+b+3}$,

[G4] $\mathrm{verr}^{\pi_{\mathfrak{m}} \otimes \xi}(\mathscr{P}'_{\mathfrak{m}}) \leq 1 - 2^{-b-1}$.

Notice that for all $\mathfrak{m} \in \mathscr{M}$ we have $\mathrm{H}(X \mid M_1 = \mathfrak{m}, X \in S) \leq n$. Hence, viewing (5), (6), (7) and (8) as upper bounds on the expected values of certain nonnegative functions of L, we may apply Markov's inequality to these four conditions and conclude that

$$\Pr[L \text{ is good}] \geq 1 - 2^{-b-3} - 2^{-b-3} - 2^{-b-3} - \frac{1 - 2^{-b}}{1 - 2^{-b-1}} \geq 2^{-b-1} - 3 \cdot 2^{-b-3} > 0\,.$$

Thus, there exists a good message. *From now on, we fix* $\mathfrak{m}$ *to be such a good message.*

We may rewrite the left-hand side of [G4] as $\mathbb{E}_{Z \sim \pi_{\mathfrak{m}}}[\Pr[\mathrm{out}(\mathscr{P}'_{\mathfrak{m}}(Z,Z)) = 0]]$. So if we define the set $T := \left\{x \in S : \Pr[\mathrm{out}(\mathscr{P}'_{\mathfrak{m}}(x,x)) = 0] \leq 1 - 2^{-b-2}\right\}$ and apply Markov's inequality again, we obtain

$$\pi_{\mathfrak{m}}(T) \geq 1 - \frac{1 - 2^{-b-1}}{1 - 2^{-b-2}} \geq 2^{-b-2}\,. \tag{9}$$

Defining the distribution $p' := \pi_{\mathfrak{m}} \mid T$ and the set $S' := \{x \in T : p'(x) \geq 2^{-n-\ell'}\}$, we now make two claims.

Claim 1: The triple (q', p', S') is an ℓ'-kernel.

Claim 2: We have $\mathrm{err}^{p' \otimes q' | S' \times S'}(\mathscr{P}'_{\mathfrak{m}}) \leq 2^{-a'}$, $\mathrm{verr}^{q' \otimes \xi | S' \times S'}(\mathscr{P}'_{\mathfrak{m}}) \leq 1 - 2^{-b'}$, and $\mathrm{icost}^{p' \otimes q'}(\mathscr{P}'_{\mathfrak{m}}) \leq c'$.

Notice that these claims essentially say that $\mathscr{P}'_{\mathfrak{m}}$ has all the properties listed in Definition 6.2, except that Bob starts $\mathscr{P}'_{\mathfrak{m}}$. Interchanging the roles of Alice and Bob in $\mathscr{P}'_{\mathfrak{m}}$ gives us the desired $[r-1, \ell', a', b', c']$-protocol, which completes the proof of the theorem.

It remains to prove the above claims. We start with Claim 1. Starting with the lower bound on $\mathrm{H}(\pi_{\mathfrak{m}})$ given by Property [G1] of the good message $\mathfrak{m}$, and using Lemma 3.8 followed by (9), we obtain

$$\mathrm{H}(p') = \mathrm{H}(\pi_{\mathfrak{m}} \mid T) \geq n - \frac{(c+\ell+1)2^{\ell+b+3} + 1}{\pi_{\mathfrak{m}}(T)} \geq n - (c+\ell+2)2^{\ell+2b+5} \geq n - \ell'\,. \tag{10}$$

We may lower bound $\mathrm{H}(q')$ using Properties [K1] and [K2] for (p, q, S) and applying Lemma 3.8. We have

$$\mathrm{H}(q') = \mathrm{H}(Y \mid Y \in S) \geq n - \frac{\ell+1}{q(S)} \geq n - 2(\ell+1) \geq n - \ell'\,.$$

Thus, (q', p', S') satisfies Property [K1] for an ℓ'-kernel. It is immediate that it also satisfies Property [K3]: by definition, for all $x \in S'$, we have $p'(x) \geq 2^{-n-\ell'}$.

It remains to verify Property [K2], which entails showing that $p'(S') \geq \frac{1}{2}$ and that $q'(S') \geq 2^{-\ell'}$. We can lower bound $p'(S')$ as follows:

$$p'(S') = 1 - \sum_{x \in \{0,1\}^n \setminus S'} p'(x) = 1 - \sum_{\substack{x \in \{0,1\}^n \\ p'(x) < 2^{-n-\ell'}}} p'(x) \geq 1 - 2^{-\ell'} \geq \frac{1}{2}\,. \tag{11}$$

To prove the second inequality, we first derive a lower bound on $\mathrm{H}(p' \mid S')$, thence on $|S'|$, and finally on $q'(S')$. We already showed that $\mathrm{H}(p') \geq n - (c+\ell+2)2^{\ell+2b+5}$, at (10). By Lemma 3.8 and (11), we get

$$\mathrm{H}(p' \mid S') \geq n - \frac{(c+\ell+2)2^{\ell+2b+5} + 1}{p'(S')} \geq n - \left((c+\ell+2)2^{\ell+2b+6} + 2\right) \geq n - (c+\ell+4)2^{\ell+2b+6}\,,$$

and so $|S'| \geq 2^{n-(c+\ell+4)2^{\ell+2b+6}}$. Since $q' = q \mid S$ and $S' \subseteq S$, we have

$$q'(S') \geq q(S') \geq |S'| \min_{y \in S'} q(y) \geq |S'| \min_{y \in S} q(y) \geq 2^{n-(c+\ell+4)2^{\ell+2b+6}} 2^{-n-\ell} = 2^{-\ell-(c+\ell+4)2^{\ell+2b+6}},$$

where the final inequality uses Property [K3]. Recalling the definition of ℓ' and applying a crude estimate (using the bound $c \geq 4$), we get $q'(S') \geq 2^{-\ell'}$. This finishes the proof of Claim 1.

We now prove Claim 2. Of the three bounds we need to prove, the verification error bound is the easiest. Recalling how T was defined, and noting that $S' \subseteq T$, we immediately obtain

$$\mathrm{verr}^{q' \otimes \xi | S' \times S'}(\mathscr{P}'_{\mathfrak{m}}) = \mathbb{E}_{Y' \sim q'}[\Pr[\mathrm{out}(\mathscr{P}'_{\mathfrak{m}}(Y', Y')) = 0] \mid Y' \in S'] \leq 1 - 2^{-b-2}.$$

To establish the overall error bound, we use

$$\begin{aligned} \mathrm{err}^{p' \otimes q' | S' \times S'}(\mathscr{P}'_{\mathfrak{m}}) &\leq \frac{\mathrm{err}^{p' \otimes q'}(\mathscr{P}'_{\mathfrak{m}})}{p'(S')q'(S')} \leq \frac{\mathrm{err}^{\pi_{\mathfrak{m}} \otimes q'}(\mathscr{P}'_{\mathfrak{m}})}{\pi_{\mathfrak{m}}(T)p'(S')q'(S')} \leq \frac{2^{-a+b+3}}{2^{-b-2} \cdot \frac{1}{2} \cdot 2^{-\ell'}} && (12) \\ &= 2^{-a+2b+6+(c+\ell)2^{\ell+2b+7}} \leq 2^{-a+(c+\ell)2^{\ell+2b+8}}, && (13) \end{aligned}$$

where the final inequality in (12) follows from Property [K2] for an ℓ'-kernel and Property [G3], and (13) just uses a crude estimate (this time $c \geq 1$ suffices). It remains to establish the information cost bound in Claim 2. We do this as follows.

$$\begin{aligned} \mathrm{icost}^{p' \otimes q'}(\mathscr{P}'_{\mathfrak{m}}) &= \mathrm{I}(XY : M_2 \ldots M_r \mid M_1 = \mathfrak{m} \wedge X \in T \wedge Y \in S) \\ &\leq \frac{\mathrm{I}(XY : M_2 \ldots M_r \mid M_1 = \mathfrak{m} \wedge (X,Y) \in S \times S) + 1}{\Pr[X \in T \mid M_1 = \mathfrak{m} \wedge (X,Y) \in S \times S]} && (14) \\ &= \frac{\mathrm{icost}^{\pi_{\mathfrak{m}} \otimes q'}(\mathscr{P}'_{\mathfrak{m}}) + 1}{\pi_{\mathfrak{m}}(T)} && (15) \\ &\leq \frac{2^{b+\ell+4}(c+1)+1}{2^{-b-2}} \leq (c+2)2^{\ell+2b+6}, && (16) \end{aligned}$$

where (14) uses Lemma 3.7, (15) uses the independence of X and Y and (16) uses Property [G2] and Eq. (9).

This completes the proof of Claim 2 and, with it, the proof of the theorem. □

The following easy corollary of Theorem 6.4 will be useful shortly.

Corollary 6.6. *Let $\tilde{n}, j, r \in \mathbb{N}$ and $a, b \in \mathbb{R}$ with $\tilde{n}$ sufficiently large, $j \geq 1$, $r \geq 1$, and $b \geq 0$. Suppose there exists an $[r, \ell, a-\ell, b, \ell]$-protocol, with $b \leq \ell = \frac{1}{8}\,\mathrm{ilog}^j \tilde{n}$. Then there exists an $[r-1, \ell', a-\ell', b+2, \ell']$-protocol with $b+2 \leq \ell' = (\mathrm{ilog}^{j-1} \tilde{n})^{1/2} \leq \frac{1}{8}\,\mathrm{ilog}^{j-1} \tilde{n}$.*

Proof. This simply boils down to the following estimation, which is valid for all sufficiently large $\tilde{n}$:

$$(\ell+\ell)2^{\ell+2b+8} = 2^7(\mathrm{ilog}^j \tilde{n})2^{(3/8)\,\mathrm{ilog}^j \tilde{n}} = 2^7(\mathrm{ilog}^{j-1} \tilde{n})^{3/8} \log(\mathrm{ilog}^{j-1} \tilde{n}) \leq (\mathrm{ilog}^{j-1} \tilde{n})^{1/2}. \qquad \square$$

6.2 Finishing the Proof

We are now ready to state and prove the main lower bound on protocols with two-sided error.

Theorem 6.7 (Restatement of Main Theorem)**.** *Let $\tilde{n} = \min\{n + \log(1-\delta), \log((1-\delta)/\varepsilon)\}$. Suppose $\delta \leq 1 - 8(\mathrm{ilog}^{r-2} \tilde{n})^{-1/8}$. Then we have $\mathrm{IC}^{\mu_u,(r)}_{\varepsilon,\delta}(\mathrm{EQ}_n) = \Omega((1-\delta)^3\,\mathrm{ilog}^{r-1} \tilde{n})$.*

Proof. We may assume that $r \le \log^* \tilde{n}$, for otherwise there is nothing to prove. The slight difference between $\tilde{n}$ above and $\hat{n}$, as in Definition 2.4, is insignificant and can be absorbed by the $\Omega(\cdot)$ notation.

Suppose, to the contrary, that there exists an r-round randomized protocol $\mathscr{P}^*$ for EQ_n, with $\mathrm{rerr}^{\mu_u}(\mathscr{P}^*) \le \varepsilon$, $\mathrm{verr}^{\mu_u}(\mathscr{P}^*) \le \delta$ and $\mathrm{icost}^{\mu_u}(\mathscr{P}^*) \le 2^{-16}(1-\delta)^3 \operatorname{ilog}^{r-1} \tilde{n}$. Recall that we denote the uniform distribution on $\{0,1\}^n$ by ξ and that $\mu_u = \xi \otimes \xi$. We have

$$\mathrm{err}^{\mu_u}(\mathscr{P}^*) = (1-2^{-n})\,\mathrm{rerr}^{\mu_u}(\mathscr{P}^*) + 2^{-n}\,\mathrm{verr}^{\mu_u}(\mathscr{P}^*) \le \varepsilon + 2^{-n}(\delta - \varepsilon) \le \varepsilon + 2^{-n}.$$

Let $\mathscr{P}^*_s$ be the private-coin protocol for EQ_n obtained from $\mathscr{P}^*$ by fixing the public random string of $\mathscr{P}^*$ to be s. We have $\mathbb{E}_s[\mathrm{err}^{\mu_u}(\mathscr{P}^*_s)] \le \varepsilon + 2^{-n}$, $\mathbb{E}_s[\mathrm{verr}^{\mu_u}(\mathscr{P}^*_s)] \le \delta$, and $\mathbb{E}_s[\mathrm{icost}(\mathscr{P}^*_s)] \le 2^{-16}(1-\delta)^3 \operatorname{ilog}^{r-1} \tilde{n}$. By Markov's inequality, there exists s such that $\mathscr{P}^*_s$ simultaneously has $\mathrm{err}^{\mu_u}(\mathscr{P}^*_s) \le 4(\varepsilon + 2^{-n})/(1-\delta)$, $\mathrm{verr}^{\mu_u}(\mathscr{P}^*_s) \le (1+\delta)/2$, and $\mathrm{icost}(\mathscr{P}^*_s) \le 2^{-14}(1-\delta)^2 \operatorname{ilog}^{r-1} \tilde{n}$: this is because

$$1 - \frac{1-\delta}{4} - \frac{2\delta}{1+\delta} - \frac{1-\delta}{4} = \frac{(1-\delta)^2}{2(1+\delta)} > 0.$$

Let $\mathscr{P} = \mathscr{P}^*_s$ for this s. Then $(\xi, \xi, \{0,1\}^n)$ is a 0-kernel and $\mathscr{P}$ is an $[r, 0, \log \frac{1-\delta}{4(\varepsilon+2^{-n})}, \log \frac{2}{1-\delta}, 2^{-14}(1-\delta)^2 \operatorname{ilog}^{r-1} \tilde{n}]$-protocol. Recalling Fact 6.3 and using $\log \frac{1-\delta}{\varepsilon+2^{-n}} \ge \tilde{n} - 1$, we see that

$$\mathscr{P} \text{ is an } \left[r, 0, \tilde{n} - 3, \log \tfrac{1}{1-\delta} + 1, 2^{-14}(1-\delta)^2 \operatorname{ilog}^{r-1} \tilde{n}\right]\text{-protocol.}$$

Put $\ell_j := \frac{1}{8} \operatorname{ilog}^j \tilde{n}$ for $j \in \mathbb{N}$. Applying round elimination (Theorem 6.4) to $\mathscr{P}$ and weakening the resulting parameters (using Fact 6.3) gives us an $[r-1, \ell_{r-1}, \tilde{n} - \ell_{r-1}, \log \frac{1}{1-\delta} + 3, \ell_{r-1}]$-protocol $\mathscr{P}'$.

The upper bound on δ gives us $\log \frac{1}{1-\delta} + 3 \le \ell_{r-1}$, and so the conditions for Corollary 6.6 apply. Starting with $\mathscr{P}'$ and applying that corollary repeatedly, each time using the looser estimate on ℓ' in that corollary, we obtain a sequence of protocols with successively fewer rounds. Eventually we reach a $[1, \ell_1, \tilde{n} - \ell_1, \log \frac{1}{1-\delta} + 2(r-1) + 1, \ell_1]$-protocol. Applying Theorem 6.4 one more time, and using the tighter estimate on ℓ' this time, we get a $[0, \tilde{n}^{1/2}, \tilde{n} - \tilde{n}^{1/2}, \log \frac{1}{1-\delta} + 2r + 1, \tilde{n}^{1/2}]$-protocol $\mathscr{Q}$. Weakening parameters again, we see that $\mathscr{Q}$ is a $[0, \tilde{n}^{1/2}, \frac{1}{2}\tilde{n}, \frac{1}{3}\log \tilde{n}, \tilde{n}^{1/2}]$-protocol. Let (p, q, S) be the $\tilde{n}^{1/2}$-kernel for $\mathscr{Q}$. By Property [K1], we have $\mathrm{H}(q) \ge n - \tilde{n}^{1/2}$. Using Lemma 3.8 and Property [K2], we then have

$$\mathrm{H}(q \mid S) \ge n - \frac{\tilde{n}^{1/2} + 1}{q(S)} \ge n - (2\tilde{n}^{1/2} + 2). \tag{17}$$

Since $\mathscr{Q}$ involves no communication, it must behave identically on any two input distributions that have the same marginal on Alice's input. In particular, this gives us the following crucial equation:

$$\Pr_{X \sim p}[\mathrm{out}(\mathscr{Q}(X,X)) = 1 \mid X \in S] = \Pr_{(X,Y) \sim p \otimes q}[\mathrm{out}(\mathscr{Q}(X,Y)) = 1 \mid (X,Y) \in S \times S]. \tag{18}$$

Let α denote the above probability. Considering the left-hand side of (18), we have

$$\alpha = 1 - \mathrm{verr}^{p \otimes \xi | S \times S}(\mathscr{Q}) \ge 2^{-\frac{1}{3}\log \tilde{n}} = \tilde{n}^{-1/3}. \tag{19}$$

On the other hand, whenever $\mathscr{Q}$ outputs 1 on an input (x,y), then either $x = y$ or $\mathscr{Q}$ errs on (x,y). Therefore, considering the right-hand side of (18), we have

$$\begin{aligned}
\alpha &\le \Pr_{(X,Y) \sim p \otimes q}[X = Y \mid (X,Y) \in S \times S] + \Pr_{(X,Y) \sim p \otimes q}[\mathrm{out}(\mathscr{P}(X,Y)) \ne \mathrm{EQ}_n(X,Y) \mid (X,Y) \in S \times S] \\
&\le \max_{x \in S} \Pr_{Y \sim q|S}[Y = x] + \mathrm{err}^{p \otimes q | S \times S}(\mathscr{Q}) \\
&\le \frac{2\tilde{n}^{1/2} + 3}{n} + 2^{-\frac{1}{2}\tilde{n}} && (20) \\
&\le 2\tilde{n}^{-1/2} + 3\tilde{n}^{-1} + 2^{-\frac{1}{2}\tilde{n}}, && (21)
\end{aligned}$$

where (20) follows from (17) by applying Lemma 3.9, and (21) uses $\tilde{n} \le n$.

The bounds (19) and (21) are in contradiction for sufficiently large $\tilde{n}$, which completes the proof. □

7 Applications of the Main Theorem

7.1 Lower Bounds for Or-Equality and Disjointness

In this section we apply our new understanding of the bounded-round information complexity of EQUALITY to obtain two other lower bounds: one for OR-EQUALITY, and the other for the much-studied DISJOINTNESS problem with small-sized sets. As we shall see, both lower bounds are tight in certain error regimes.

Theorem 7.1 (Lower Bound for Or-Equality)**.** *Let $k,n,r \in \mathbb{N}$ and $\delta,\varepsilon \in [0,1]$. Put $\varepsilon' = \varepsilon + k/2^n$ and $\tilde{n} = \log \frac{1-\delta}{\varepsilon'}$. For $\delta < 1 - 8(\operatorname{ilog}^{r-2}\tilde{n})^{-1/8}$, we have*

$$\mathrm{R}^{(r)}_{\varepsilon,\delta}(\mathrm{OREQ}_{n,k}) \geq k \cdot \mathrm{IC}^{\mu,(r)}_{\varepsilon',\delta}(\mathrm{EQ}_n) = \Omega(k(1-\delta)^3 \operatorname{ilog}^{r-1}\tilde{n})\,.$$

Proof. We just need to show the first inequality and then apply Theorem 2.6. That inequality is proved via standard direct sum arguments for information complexity [17, 4, 5]. In fact, the old simultaneous-message lower bound for $\mathrm{OREQ}_{n,k}$ from Chakrabarti et al. [17] applies more-or-less unchanged. For completeness, we now give a self-contained proof.

Let $\mathscr{P}$ be an r-round protocol for $\mathrm{OREQ}_{n,k}$ with $\mathrm{rerr}(\mathscr{P}) \leq \varepsilon$, $\mathrm{verr}(\mathscr{P}) \leq \delta$, and $\mathrm{R}^{(r)}_{\varepsilon,\delta}(\mathrm{OREQ}_{n,k}) \geq \max\{\mathrm{rcost}(\mathscr{P}), \mathrm{vcost}(\mathscr{P})\}$. Alice and Bob solve EQ_n by the following protocol $\mathscr{Q}_j$, where j is some fixed index in $\{1,2,\ldots,k\}$. Given an input $(x,y) \in \{0,1\}^n \times \{0,1\}^n$, they generate $\mathbf{X} := (X_1,\ldots,X_k) \sim \xi^{\otimes k}$ and $\mathbf{Y} := (Y_1,\ldots,Y_k) \sim \xi^{\otimes k}$ respectively, using private coins. They "plug in" x and y into the jth coordinates of $\mathbf{X}$ and $\mathbf{Y}$ respectively, thereby creating

$$\mathbf{Z}_{j,x} := (X_1,\ldots,X_{j-1},x,X_{j+1},\ldots,X_k) \text{ and } \mathbf{W}_{j,y} := (Y_1,\ldots,Y_{j-1},y,Y_{j+1},\ldots,Y_k)\,,$$

respectively. Finally, they emulate $\mathscr{P}$ on input $(\mathbf{Z}_{j,x}, \mathbf{W}_{j,y})$. Observe that

$$\mathrm{OREQ}_{n,k}(\mathbf{Z}_{j,x}, \mathbf{W}_{j,y}) \neq \mathrm{EQ}_n(x,y) \implies (x \neq y) \wedge \big(\exists\, i \in [k] \setminus \{j\} : X_i = Y_i\big)\,.$$

Therefore, $\mathrm{verr}(\mathscr{Q}_j) \leq \mathrm{verr}(\mathscr{P}) \leq \delta$ and, by a union bound,

$$\mathrm{rerr}(\mathscr{Q}_j) \leq \mathrm{rerr}(\mathscr{P}) + \sum_{i=1}^{n} \Pr[X_i = Y_i] \leq \varepsilon + k/2^n = \varepsilon'\,.$$

Since $\mathscr{Q}_j$ solves EQ_n with these error guarantees, it follows that $\mathrm{icost}^{\mu}(\mathscr{Q}_j) \geq \mathrm{IC}^{\mu,(r)}_{\varepsilon',\delta}(\mathrm{EQ}_n)$.

Now, let $(X,Y) \sim \mu$ and let $\mathfrak{R}$ denote the public randomness used by $\mathscr{P}$. We can now lower bound $\mathrm{R}^{(r)}_{\varepsilon,\delta}(\mathrm{OREQ}_{n,k})$ as follows:

$$\begin{aligned}
\mathrm{R}^{(r)}_{\varepsilon,\delta}(\mathrm{OREQ}_{n,k}) &\geq \max\nolimits_{x_1,\ldots,x_k,y_1,\ldots,y_k \in \{0,1\}^{kn} \times \{0,1\}^{kn}} \mathrm{cost}(\mathscr{P}; x_1,\ldots,x_k,y_1,\ldots,y_k) \\
&\geq \mathbb{E}[\mathrm{cost}(\mathscr{P}; X_1,\ldots,X_k,Y_1,\ldots,Y_k)] \\
&\geq \mathrm{H}(\mathscr{P}(X_1,\ldots,X_k,Y_1,\ldots,Y_k)) \qquad (22) \\
&\geq \mathrm{I}(\mathscr{P}(X_1,\ldots,X_k,Y_1,\ldots,Y_k) : X_1Y_1\ldots X_kY_k \mid \mathfrak{R}) \\
&\geq \sum_{j=1}^{k} \mathrm{I}(\mathscr{P}(X_1,\ldots,X_k,Y_1,\ldots,Y_k) : X_i,Y_i \mid \mathfrak{R}) \qquad (23) \\
&= \sum_{j=1}^{k} \mathrm{I}(\mathscr{Q}_j(X,Y) : XY \mid \mathfrak{R}) \qquad (24) \\
&= \sum_{j=1}^{k} \mathrm{icost}^{\mu}(\mathscr{Q}_j) \geq k \cdot \mathrm{IC}^{\mu,(r)}_{\varepsilon',\delta}(\mathrm{EQ}_n)\,,
\end{aligned}$$

where (22) uses Fact 3.6 and (23) uses the independence of $\{X_1Y_1,\dots,X_kY_k\}$ and the resulting subadditivity of mutual information, and (24) holds because, for all $j\in[k]$, the distributions of $(\mathscr{Q}_j(X,Y),X,Y,\mathfrak{R})$ and $(\mathscr{P}(X_1,\dots,X_k,Y_1,\dots,Y_k),X_j,Y_j,\mathfrak{R})$ are identical. This completes the proof. □

By plugging in $\varepsilon=0$, $\delta=0$ in Theorem 7.1 we obtain the following corollary.

Corollary 7.2. $\mathrm{R}^{(r)}_{0,0}(\textsc{oreq}_{n,k})=\Omega(k\operatorname{ilog}^{r-1}(n-\log k))$. □

Armed with the above lower bound, we now derive a lower bound for k-DISJ via a simple reduction, which is probably folklore. For completeness, we again give a formal proof. A similar observation has also been made by Sağlam and Tardos [49]. Note that the reduction interchanges verification and refutation errors.

Lemma 7.3 (Reductions from OREQ to k-DISJ and from $\textsc{eq}^k$ to k-$\textsc{int}_N$)**.** *Let k,N be integers such that $N\geq k^c$ for some constant $c>2$. Let $n=\left\lfloor\log\left(\frac{N}{k}\right)\right\rfloor$. If there exists a protocol $\mathscr{P}$ for k-$\textsc{disj}_N$ then there exists a protocol $\mathscr{Q}$ for $\textsc{oreq}_{n,k}$ such that $\operatorname{rerr}(\mathscr{Q})\leq\operatorname{verr}(\mathscr{P})$ and $\operatorname{verr}(\mathscr{Q})\leq\operatorname{rerr}(\mathscr{P})$ and $\operatorname{vcost}(\mathscr{Q})\leq\operatorname{rcost}(\mathscr{P})$ and $\operatorname{rcost}(\mathscr{Q})\leq\operatorname{vcost}(\mathscr{P})$. Moreover, the same reduction can be applied between $\textsc{eq}^k_n$ and k-$\textsc{int}_N$.*

Proof. Given an input instance $(x_1,\dots,x_k,y_1,\dots,y_k)$ of $\textsc{oreq}_{n,k}$, we can transform it into an instance (A,B) of k-$\textsc{disj}_N$ as follows:

$$A=\{x_1,\,x_2+2^n,\,x_3+2\cdot2^n,\,\dots,\,x_k+(k-1)2^n\}$$
$$B=\{y_1,\,y_2+2^n,\,y_3+2\cdot2^n,\,\dots,\,y_k+(k-1)2^n\}.$$

It is easy to observe that $A\cap B\neq\emptyset$ iff $\exists\, i\in[k]$ such that $x_i=y_i$ because $x_i\in\{0,1,\dots,2^n-1\}$. Therefore, $\textsc{oreq}_{n,k}(x_1,\dots,x_k,y_1,\dots,y_k)=\neg k\text{-}\textsc{disj}_N(A,B)$, which completes the proof. The reduction from $\textsc{eq}^k_n$ to k-$\textsc{int}_N$ is the same. □

Corollary 7.4. *We have:*

$$\mathrm{R}^{(r)}_{\delta,\varepsilon}(k\text{-}\textsc{disj}_N)\geq\mathrm{R}^{(r)}_{\varepsilon,\delta}(\textsc{oreq}_{\lfloor\log(N/k)\rfloor,k})$$
$$\mathrm{R}^{(r)}_{\delta,\varepsilon}(k\text{-}\textsc{int}_N)\geq\mathrm{R}^{(r)}_{\varepsilon,\delta}(\textsc{eq}^k_{\lfloor\log(N/k)\rfloor}).$$

□

Combining Corollary 7.4 with Theorem 7.1, we arrive at the following theorem.

Theorem 7.5 (Lower Bound for k-Disjointness)**.** *Let $k,N,r\in\mathbb{N}$, $\varepsilon,\delta\in[0,1]$ and $c>2$ be such that $N\geq k^c$ and $\delta<1-8(\operatorname{ilog}^{r-2}\tilde{n})^{-1/8}$, where $\tilde{n}=\log\frac{1-\delta}{\varepsilon+k^2/N}$. Then*

$$\mathrm{R}^{(r)}_{\delta,\varepsilon}(k\text{-}\textsc{disj}_N)=\Omega(k(1-\delta)^3\operatorname{ilog}^{r-1}\tilde{n}).$$

In particular, with $\delta=1-\Omega(1)$ and $\varepsilon\leq k^{-\Theta(1)}$, we have $\mathrm{R}^{(r)}_{\delta,\varepsilon}(k\text{-}\textsc{disj}_N)=\Omega(k\operatorname{ilog}^r k)$. □

By plugging in $\varepsilon=\delta=0$ above we arrive at a further special case that is worth highlighting.

Corollary 7.6. *With $N\geq k^{2+\Omega(1)}$, we have $\mathrm{R}^{(r)}_{0,0}(k\text{-}\textsc{disj}_N)=\Omega(k\operatorname{ilog}^r k)$.* □

7.2 Tightness

Our lower bounds in Section 7.1 have the weakness that they apply only in zero-error or small-error settings. However, they are still tight in the following sense. We can design protocols that give matching *upper* bounds under similarly small error settings. For OREQ, we give such a protocol below. For k-DISJ, a suitable analysis of a recent protocol of Sağlam and Tardos [49] gives similar results.

Theorem 7.7. *For all $r < \log^* k$, there exists a r-round protocol $\mathscr{P}$ for $\text{OREQ}_{n,k}$ with worst-case communication cost $O(k \operatorname{ilog}^r k)$, $\text{rerr}(\mathscr{P}) < 2^{-\prod_{j=1}^{r} \operatorname{ilog}^j k}$, and $\text{verr}(\mathscr{P}) = 0$.*

Proof. We begin with a high-level sketch of the proof, before giving formal proof details. Alice begins the protocol by sending, in parallel, k different t-bit equality tests, one for each of her inputs. Note that for any i where $x_i \neq y_i$, Bob witnesses non-equality with probability $1 - 2^{-t}$. Assuming $\text{OREQ}_{n,k}(x,y) = 0$, there will be roughly $k/2^t$ coordinates i where $x_i \neq y_i$ has not yet been witnessed. Bob now tells Alice which of his coordinates remain "alive" and sends t'-bit equality tests for each of *these* coordinates, where $t' = 2^t$. Note that Bob's overall communication is roughly k bits, and that after receiving this message, Alice witnesses non-equality on all but a $2^{-t'}$-fraction of unequal pairs. In each round, players end up sending an exponentially longer equality test on an exponentially smaller number of coordinates. When communication ends, players output $\text{OREQ}(x_1, \ldots, x_k, y_1, \ldots, y_k) = 1$ unless $x_i \neq y_i$ has been witnessed for all i. One potential issue with the above protocol is that too many coordinates could remain, and players wouldn't be able to communicate exponentially more bits about the remaining coordinates. This could happen both when an unusually large number of equality tests fail, or just for the simple reason that $x_i = y_i$ for many coordinates. In either case, the players simply abort and output $\text{OREQ}_{n,k} = 1$. This will cause an increase in error, but the increase will be small, and it will only increase the false positive rate. A formal proof lies below.

Before formally analyzing the complete protocol, we introduce some additional terminology and notation. For $0 \leq j \leq r$, let $z_j := \operatorname{ilog}^{r-j} k$ and $\delta_j := 1/z_j$. For $1 \leq j < r$, let $t_j := 2z_{j-1}$, and let $t_r := 2\prod_{j=1}^{r} \operatorname{ilog}^j k$. Finally, let $c_1 := 2k$ and for $2 \leq j \leq r$, let $c_j := 2k \prod_{i=1}^{j-1} \delta_i$. Note that $t_r = (4k \operatorname{ilog}^r k)/c_r$.

Now we are ready to formally describe our protocol. The protocol proceeds in a number of rounds. Throughout, players maintain a vector $w \in \{0,1\}^k$ (initialized to $w = 1^k$), where $w_i = 0$ iff $x_i \neq y_i$ has been witnessed. Coordinate i is deemed "live" if $w_i = 1$. Each round of communication is a three part message—first, a bit indicating whether to abort the protocol; second, an updated description of which coordinates remain live, and finally an equality test for each remaining live coordinate. Say coordinate i is live after j rounds if $x_i \neq y_i$ has not been witnessed by the first j rounds of equality tests. Note that the player that *receives* the jth message that determines which coordinates are live after j rounds. The sender of the jth message must wait until round $j+2$ to learn which coordinates failed the jth equality test. We describe this more completely below.

In the first round of communication, Alice sends a t_1-bit equality test for each of the k live coordiantes, at a total cost of $kt_1 = 2kz_0 = O(k \operatorname{ilog}^r k)$ bits. Assuming the protocol has not yet aborted, in the jth round of communication ($1 < j \leq r$), the player to speak first updates her copy of w by considering the $(j-1)$th message: first, she notes which i were live at the end of round $j-2$ using the second part of the $(j-1)$th message. Then, for each live i, she sets uses the third part of the message to test equality on (x_i, y_i), setting $w_i = 0$ if $x_i \neq y_i$ has been witnessed. At this point, w describes the set of coordinates that are live after $j-1$ rounds. Now, if more than c_j coordinates remain live, she sends "1", signifying that the protocol should abort and output $\text{OREQ}_{n,k} = 1$. Otherwise, she sends 0, followed by a description of which coordinates remain live, followed by a t_j-bit equality test for each of the remaining live coordinates. In this way, the jth message is at most $O(1 + k + c_j t_j)$ bits.

The receiver of the final message updates his copy of w, evaluates each equality test, and outputs $\text{OREQ}_{n,k} = 1$ if any coordinates remain live. Otherwise, he outputs $\text{OREQ}_{n,k} = 0$.

The overall communication is $O(kr + \sum_{j=1}^{r} c_j t_j)$. Note that $c_1 t_1 = 4k \operatorname{ilog}^r k$, and $c_r t_r = 4k \operatorname{ilog}^r k$. Furthermore, since $z_j > 2$ for all $j \geq 1$, we have for all $2 \leq j < r$

$$c_j t_j = (2k \prod_{i=1}^{j-1} \delta_i) \cdot (2z_{j-1}) = 4k \prod_{i=1}^{j-2} \delta_i = c_{j-1} t_{j-1} \delta_{j-2} < \frac{c_{j-1} t_{j-1}}{2} .$$

Thus, the summation $\sum_{j=1}^{r-1} c_j t_j$ telescopes, and the overall communication is $O(kr + k \operatorname{ilog}^r k) = O(k \operatorname{ilog}^r k)$.[7] Note also that the protocol outputs $\text{OREQ}_{n,k} = 0$ only when $x_i \neq y_i$ was witnessed for every i. Thus, the protocol produces no false negatives.

A false positive can happen for one of two reasons: either the protocol aborts (outputting $\text{OREQ}_{n,k} = 1$), or one or more coordinates remain live at the end of the protocol, despite having $x_i \neq y_i$ for all i.

In the former case, note that (conditioned on not aborting before round j) we have at most c_j live coordinates during round j. Players execute a t_j-bit equality test during this round. Thus, a coordinate remains live after this test with probability at most $2^{-t_j} = 2^{-2z_{j-1}} = \delta_j^2 < \delta_j/2$. By a Chernoff bound and the fact that $c_{j+1} = c_j \delta_j$, the probability that more than c_{j+1} coordinates remain live after round j is at most $e^{-c_j \delta_j^2/8} < e^{-k^{1-\varepsilon}}$ for any $\varepsilon > 0$ and large enough k. In the latter case, note that the final equality test uses $t_r = 2 \prod_{j=1}^{r} \operatorname{ilog}^j k$ bits. Therefore, players *fail* to witness $x_i \neq y_i$ with probability at most $2^{-t_r} = 2^{-2 \prod_{j=1}^{r} \operatorname{ilog}^j k}$. By a union bound, the overall false positive rate is at most $2^{-\prod_{j=1}^{r} \operatorname{ilog}^j k}$. □

7.3 Private Intersection and Strong Direct Sum for Equality

We now prove our result for PRIVATE-INTERSECTION.

Theorem 7.8 (Lower Bound for PRIVATE-INTERSECTION)**.** *Let $k, N, r \in \mathbb{N}$ and $c > 2$ be such that $N \geq k^c$. Then:*

$$\mathrm{R}^{(r)}_{1/3,1/3}(k\text{-INT}_N) = \Omega(k \operatorname{ilog}^r k) .$$

Using the reduction from Corollary 7.4 it suffices to show the lower bound for EQ_n^k, where $n = \lfloor \log(N/k) \rfloor$. In the proof we will use the following modification of the strong direct sum theorem of [42] (Theorem 2.1), which uses protocols with abortion (see definitions in Section 3.2). The simulation procedure used in the proof of this theorem in [42] preserves the number of rounds in the protocol, which allows us to state their theorem as:

Theorem 7.9 (Strong Direct Sum [42])**.** *Let $\delta \leq 1/3$. Then for every function $f \colon \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ and distribution λ on $\mathcal{X} \times \mathcal{Y} \times \mathrm{D}$ with marginal μ_p on $\mathcal{X} \times \mathcal{Y}$ and marginal ν_p on D, such that μ_p is partitioned by ν_p, it holds that $\mathrm{IC}^{\mu_p^k,(r)}_{\delta}(f^k | \nu_p^k) \geq \Omega(k) \, \mathrm{IC}^{\mu_p,(r)}_{\frac{1}{20},\frac{1}{10},\frac{\delta}{k}}(f | \nu_p)$.*

Using the direct sum above it remains to show the following:

Lemma 7.10. *There exists a distribution on $\mathcal{X} \times \mathcal{Y} \times \mathcal{D}$ with marginals μ_p on $\mathcal{X} \times \mathcal{Y}$ and ν_p on $\mathcal{D}$, such that ν_p partitions μ_p and*

$$\mathrm{IC}^{\mu_p,(r)}_{1/20,1/10,\delta/k}(\mathrm{EQ}_n | \nu_p) = \Omega(\operatorname{ilog}^r k) .$$

Proof. In the proof we can use the same hard distribution as in [42]. To construct μ_p and ν_p, let D_0 be a random variable uniformly distributed on $\{0,1\}$ and let $\mathbf{D}$ be a random variable uniformly distributed on $\{0,1\}^n$. Let $(\mathbf{X}, \mathbf{Y})$ be a random variable supported on $\{0,1\}^n \times \{0,1\}^n$ such that, conditioned on $D_0 = 0$

[7] For some values of k, r, we might have $r > \operatorname{ilog}^r k$. In fact, it is possible to describe the set of c_j live coordinates using $log(\binom{k}{c_j})$ bits. This sum also telescopes, so it is possible to reduce the $O(kr)$ cost of describing $\{c_j\}$ to just $O(k)$ bits. Thus, the overall cost remains $O(k \operatorname{ilog}^r k)$.

we have $\mathbf{X}$ and $\mathbf{Y}$ distributed independently and uniformly on $\{0,1\}^n$, and conditioned on $D_0 = 1$ we have $\mathbf{X} = \mathbf{Y} = \mathbf{D}$. Let μ_p be the distribution of $(\mathbf{X},\mathbf{Y})$ and let ν_p be the distribution of $(D_0\mathbf{D})$. Note that ν_p partitions μ_p. Also, this distribution satisfies that $\Pr[\mathbf{X} = \mathbf{Y}] \geq 1/3$ and $\Pr[\mathbf{X} \neq \mathbf{Y}] \geq 1/3$.

Let W be a random variable distributed according to ν_p. Let E be an indicator variable over the private randomness of $\mathscr{P}$ which is equal to 1 if and only if conditioned on this private randomness $\mathscr{P}$ satisfies that it aborts with probability at most $1/10$ and succeeds with probability at least $1-\delta/k$ conditioned on non-abortion. Given such protocol with abortion $\mathscr{P}$ we transform it into a protocol $\mathscr{P}'$ which never aborts, has almost the same information complexity and gives correct output on non-equal instances with high probability, while being correct on equal instances with constant probability. This is done by constructing $\mathscr{P}'$ so that whenever $\mathscr{P}$ outputs "abort", the output of $\mathscr{P}'$ is $X \neq Y$, otherwise $\mathscr{P} = \mathscr{P}'$. Under the distribution μ_p conditioned on the event $E = 1$ the protocol $\mathscr{P}'$ has the property that if $X \neq Y$, then it outputs $X = Y$ with probability at most $(1/k)/\Pr_{\mu_p}[X \neq Y] \leq 3/k$. However, if $X = Y$, then the protocol may output $X \neq Y$ with probability $1/10 + (1/k)/\Pr_{\mu'_p}[X = Y] \leq 1/10 + 3/k \leq 1/5$, where the latter follows for $k \geq 30$. Thus, conditioned on $E = 1$, the protocol $\mathscr{P}'$ has failure probability $\varepsilon = 1/k$ on non-equal instances $X \neq Y$, and constant failure probability $\delta = 1/5$ on equal instances $X = Y$, as desired. In this regime we can use Theorem 2.6.

We have:

$$\begin{aligned}\mathrm{IC}^{\mu_p,(r)}_{1/20,1/10,\delta/k}(\mathrm{EQ}_n|\nu_p) &\geq \mathrm{I}(\mathscr{P} : X,Y|W) \\ &= \Omega(\mathrm{I}(\mathscr{P} : X,Y|W,E=1)) - 1 \\ &= \Omega(\mathrm{I}(\mathscr{P}' : X,Y|W,E=1)) - 2.\end{aligned}$$

Here the inequality is by definition of information compelxity and the equalities follows from Fact 3.4 together with the fact that $H(E) \leq 1$, $\Pr[E=1] = 19/20$, and the fact that the transcripts of the protocols $\mathscr{P}$ and $\mathscr{P}'$ only differ in a single bit. The right-hand side can be bounded as follows.

Proposition 7.11.

$$\mathrm{I}(\mathscr{P}' : X,Y|W,E=1)) = \Omega(\mathrm{IC}^{\mu,(r)}_{1/k,1/5}(\mathrm{EQ}_n)).$$

Proof. This follows from the construction of the distributions μ_p and ν_p that we use. If $D_0 = 0$ then $\mathbf{X} = \mathbf{Y}$ and the information revealed by $\mathscr{P}$ is equal to zero. Otherwise, if $D_0 = 1$ then the distribution of $(\mathbf{X},\mathbf{Y})$ is uniform. Because the latter happens with probability $1/2$ we have $\mathrm{I}(\mathscr{P}' : X,Y|W,E=1)) \geq 1/2 \cdot \mathrm{IC}^{\mu,(r)}_{1/k,1/5}(\mathrm{EQ}_n))$ as desired. □

Using Proposition 7.11 we have $\mathrm{IC}^{\mu_p,(r)}_{1/20,1/10,\delta/k}(\mathrm{EQ}_n|\nu_p) = \Omega(\mathrm{IC}^{\mu,(r)}_{1/k,1/5}(\mathrm{EQ}_n))$. The proof is completed by noting that setting $\varepsilon = 1/k$ and $\delta = 1/5$ in Theorem 2.6 gives $\mathrm{IC}^{\mu,(r)}_{1/k,1,5}(\mathrm{EQ}_n) = \Omega(\mathrm{ilog}^r k)$. □

8 Two-Party Set Intersection

In this section we give upper bounds in both private and public randomness model. In the private random string model, the players do not share a random string, but rather are allowed to use private randomness. By a result of Newman [45], any problem that can be solved in the public random string model can be solved in the private random string model, adding only $O(\log\log T)$ to the communication complexity, where T is the number of different inputs to the players. One unfortunate aspect of this reduction is that it is non-constructive in the sense that for each input length n, the protocol either uses a hard-wired advice string that depends on n, or the players must search for the advice string, which doesn't require communication but can result in unnecessary computation. We give our upper bounds in the public random string model, but

describe how to translate them into constructive protocols in the private random string model, preserving optimality.

We start by describing a simple protocol with linear communication in Section 8.1 and then show how to achieve an optimum round vs. communication trade-off in Section 8.2 and Section 8.3.

8.1 Warmup: An $O(\sqrt{k})$-Round Protocol

Theorem 8.1. *There exists an $O(\sqrt{k})$-round constructive randomized protocol for k-INT_N with success probability $1-1/\text{poly}(k)$. In the model of shared randomness the total expected communication is $O(k)$ and in the model of private randomness it is $O(k+\log\log N)$*

Proof. W.l.o.g we can assume that $N = k^c$ for a constant $c > 2$ since if the universe size is $N > k^c$ then parties can pick a random hash function $H\colon [N] \to [k^c]$, which gives no collisions on the elements in $S \cup T$ with probability at least $1 - 1/\Omega(k^{c-2})$.

The parties pick a random hash function $h\colon [N] \to [k]$. For a set $U \subseteq [N]$ we use notation $U_i = h^{-1}(i) \cap U$ for the preimage of i in U. Using preimages S_i and T_i the parties construct a collection of instances of EQUALITY, which contains an instance of $\text{EQUALITY}(s,t)$ for every $(s,t) \in S_i \times T_i$ for every $i \in [k]$.

Formally, for two sets of instances of a communication problem C, say $C_1 = C(x_1,y_1),\ldots,C(x_i,y_i)$ and $C_2 = C(x'_1,y'_1),\ldots,C(x'_j,y'_j)$ let's denote their concatenation, which corresponds to solving C_1 and C_2 simultaneously as

$$C_1 \sqcup C_2 = (x_1,y_1),\ldots,(x_i,y_i),(x'_1,y'_1),\ldots(x'_j,y'_j).$$

Let's denote as $E_i = \bigsqcup_{(s,t)\in(S_i\times T_i)} \text{EQ}(s,t)$ the collection of instances of equality corresponding to hash value i. The collection of all instances constructed by the parties is $E = \bigsqcup_{i=1}^{k} E_i$.

The expected number of instances $\mathbb{E}[|E|]$ is given as:

$$\begin{aligned}\mathbb{E}[|E|] &= \mathbb{E}\left[\sum_{i=1}^{k} |S_i||T_i|\right] = \sum_{i=1}^{k} \mathbb{E}[|S_i||T_i|] \\ &\leq \sum_{i=1}^{k} \mathbb{E}[|(S\cup T)_i|^2] = \sum_{i=1}^{k} Var[|(S\cup T)_i|] + \mathbb{E}[|(S\cup T)_i|]^2 \end{aligned} \tag{25}$$

Given that for a set Z, the random variable $|Z_i|$ is distributed according to a binomial distribution $B(|Z|,1/k)$, for each i we have $Var[|(S\cup T)_i|] \leq 2k\cdot(1/k)(1-1/k) \leq 2$ and $\mathbb{E}[|(S\cup T)_i|] \leq 2$ so $\mathbb{E}[|E|] \leq 6k$.

We use the following result of [23]:

Theorem 8.2 ([23])**.** *There exists a constructive randomized protocol for EQ_n^k with $O(\sqrt{k})$ rounds, which has success probability $2^{-\Omega(\sqrt{k})}$. In the public randomness model the expected total communication is $O(k)$ and in the private randomness model it is $O(k+\log n)$.*

In the shared randomness model the result now follows immediately. In the private randomness model the parties need to construct two random hash functions H and h, using Fact 3.1 with only $O(\log N) + O(\log k) = O(\log N)$ random bits. These bits are exchanged through the channel in the first round of the protocol and are added to the total communication, bringing it down to $O(k+\log N)$. To further reduce the communication we can use the hashing scheme of Fredman, Komlos and Szemeredi [24] as the first step of the protocol. In [24] it is shown that mapping elements $[N]$ by taking a remainder modulo a random prime $q = \tilde{O}(k^2\log n)$ gives no collisions on a subset of size $O(k)$ with probability $1-1/\text{poly}(k)$. Applying this result to $S\cup T$ we can reduce the length of strings in the instances of equality down to $O(\log k+\log\log N)$. Thus, we can now specify the pairwise independent hash function using only $O(\log k+\log\log N)$ random bits. See Appendix A.1.1 in [34] for a detailed discussion.

□

8.2 Auxiliary Protocols

We first describe auxiliary protocols BASIC-INTERSECTION (Lemma 8.3) and EQUALITY (Fact 8.5) that we use as building blocks in our main algorithm in Section 8.3. For a two-party communication protocol $\mathscr{P}$ we denote the output of the protocol for the first party as $\mathscr{P}_A(x,y)$ and for the second party as $\mathscr{P}_B(x,y)$.

Lemma 8.3 (Protocol BASIC-INTERSECTION(S,T))**.** *For any integer $i \geq 1$, there exists a public-coin protocol $\mathscr{P}$ such that for any $S,T \subset [n]$, the sets $S' = \mathscr{P}_A(S,T)$ and $T' = \mathscr{P}_B(S,T)$ satisfy the following properties:*

1. *$S' \subseteq S, T' \subseteq T$.*
2. *If $S \cap T = \emptyset$ then $S' \cap T' = \emptyset$ with probability 1.*
3. *If $S \cap T \neq \emptyset$ then $(S \cap T) \subseteq (S' \cap T')$. Also, with probability $1 - 1/N^i$ it holds that $S' = T' = (S \cap T)$.*

The total communication in the protocol is

$$O\left(i \cdot (|S| + |T|) \log(|S| + |T|)\right)$$

and the protocol can be executed in 4 rounds.

Note that Lemma 8.3 guarantees that $S' \cap T'$ is always a superset of the intersection. Also, if the sets S' and T' are equal then each of them is exactly the intersection of S and T.

Proof. The parties first exchange the sizes of their sets $|S|$ and $|T|$ and determine $m = |S| + |T|$. Using shared randomness they pick a random hash function $h\colon [n] \to [t]$, where $t = \Theta(m^{i+2})$. They exchange sets $h(S)$ and $h(T)$ using total communication $O(i \cdot m \log m)$. The outcome of the protocol is $\mathscr{P}_A(S,T) = h^{-1}(h(T)) \cap S$ and $\mathscr{P}_B(S,T) = h^{-1}(h(S)) \cap T$. Since exchanging the sizes of the sets takes two rounds and another two rounds are required to exchange $h(S)$ and $h(T)$, the total number of rounds of communication is 4.

By construction we have $S' = h^{-1}(h(T)) \cap S \subseteq S$ and similarly $T' \subseteq T$ so the first property holds. If $S \cap T = \emptyset$ then $S' \cap T' = (h^{-1}(h(T)) \cap S) \cap (h^{-1}(h(S)) \cap T) \subseteq (S \cap T) = \emptyset$ and the second property holds. Because $S \subseteq h^{-1}(h(S))$ and $T \subseteq h^{-1}(h(T))$ we have

$$S \cap T \subseteq (h^{-1}(h(T)) \cap S) \cap (h^{-1}(h(S)) \cap T) = S' \cap T',$$

the first part of the third property. Moreover, if the hash function h has no collisions among $S \cup T$ then

$$S' = h^{-1}(h(T)) \cap S = T \cap S$$

and

$$T' = h^{-1}(h(S)) \cap T = S \cap T.$$

The proof is completed using the analysis of collision probability given by Fact 3.1. □

We have the following corollary.

Corollary 8.4. *If for the protocol $\mathscr{P}$ in Lemma 8.3 it holds that $\mathscr{P}_A(S,T) = \mathscr{P}_B(S,T)$ then*

$$\mathscr{P}_A(S,T) = \mathscr{P}_B(S,T) = S \cap T.$$

In our main protocol in Section 8.3 we will use an EQ_n test with the following guarantees to verify correctness of the protocol BASIC-INTERSECTION. The following guarantee is achieved by a protocol, which uses a random hash function h into k bits.

Fact 8.5. There exists a public-coin protocol $\mathscr{P}$ for EQ_n with the following properties.

1. If $x = y$ then $\mathscr{P}_A(x,y) = \mathscr{P}_B(x,y) = 1$ with probability 1.
2. If $x \neq y$ then $\mathscr{P}_A(x,y) = \mathscr{P}_B(x,y) = 0$ with probability at least $1 - 1/2^k$.

The total communication in the protocol is $O(k)$ and it can be executed in two rounds.

8.3 The Main Protocol

In this section we give the full protocol, proving Theorem 2.8.

Proof. For $r = 1$ the parties use shared randomness to pick a hash function $h\colon [N] \to [k^c]$ for $c > 2$. Then each of the parties uses $ck \log k$ bits to exchange $h(S)$ and $h(T)$ respectively. By Fact 3.1 the probability that h has a collision on a set $S \cup T$ is at most $1 - 1/\Theta(k^{c-2})$.

For $r > 1$ consider a tree $\mathscr{T}$ of depth r with the set of nodes at the i-th level for $0 \le i \le r$ denoted as L_i (these are the nodes at distance i from the leaves). Let the degree at the i-th level for $2 \le i \le r$ be equal to $d_i = \operatorname{ilog}^{r-i} k / \operatorname{ilog}^{r-i+1} k$ and the degree at the first level is $d_1 = \operatorname{ilog}^{r-1} k$. Note that this guarantees that the total number of leaves in the tree is k. For a node $v \in \mathscr{T}$, let $c(v)$ denote the set of children of v. For a node $v \in \mathscr{T}$, let $\mathscr{C}(v)$ denote the set of all leaves in the subtree of v. Note that for a node $v \in L_i$ the number of such leaves is $|\mathscr{C}(v)| = \operatorname{ilog}^{r-i} k$.

Definition 8.6 (Set assignment)**.** A set assignment $\mathscr{A}$ to the leaves of $\mathscr{T}$ is a vector $\mathscr{A} = (\mathscr{A}_1, \ldots, \mathscr{A}_k)$, consisting of k sets. We say that the set $\mathscr{A}_\ell$ is assigned to a corresponding leaf ℓ in $\mathscr{T}$.

Every set assignment to the leaves of $\mathscr{T}$ naturally induces a set assignment on all internal vertices of $\mathscr{T}$. Let $\mathscr{A} = (\mathscr{A}_1, \ldots, \mathscr{A}_k)$ be a set assignment for the leaves of $\mathscr{T}$. For every internal node $v \in \mathscr{T}$ we denote an assignment induced at this vertex by $\mathscr{A}$ as $\mathscr{A}_v = \cup_{i \in \mathscr{C}(v)} \mathscr{A}_i$.

Now we describe the protocol used by the parties. First, Alice and Bob use shared randomness to pick a hash function $h\colon [N] \to [k]$. Using this hash function they define initial assignments of sets S^{-1} and T^{-1} respectively as follows. For a leaf $\ell \in [k]$ of $\mathscr{T}$, let $S^{-1}_\ell = h^{-1}(h(\ell)) \cap S$ and $T^{-1}_\ell = h^{-1}(h(\ell)) \cap T$.

Then the protocol proceeds in r stages. In stage i for $0 \le i < r$ the parties construct new assignments to the leaves of $\mathscr{T}$, which induce new assignments on the internal nodes. We will show that after r stages the parties obtain an assignment to the leaves, such that with high probability the set induced by this assignment in the root of $\mathscr{T}$ is exactly $S \cap T$. We use notation S^i and T^i respectively for the i-th assignment that the parties make to the leaves of the tree. The description of the i-th stage is given as Algorithm 2. This completes the description of the protocol.

Algorithm 2: Protocol for k-INT$_N$. Round i.

Input: Sets $S, T \in [k]^k$, assignments S^{i-1}, T^{i-1}.

1: For every $v \in L_i$ run the protocol EQUALITY(S^{i-1}_v, T^{i-1}_v) with success probability $1 - 1/(\operatorname{ilog}^{r-i-1} k)^4$.
2: Let $\mathscr{F}$ be the set of vertices for which the equality protocol above returns $S^{i-1}_v \neq T^{i-1}_v$. We call these vertices *failed*.
3: For every $v \in \mathscr{F}$ and every leaf $u \in \mathscr{C}(v)$ run BASIC-INTERSECTION(S^{i-1}_u, T^{i-1}_u) with success probability $1 - 1/(\operatorname{ilog}^{r-i-1} k)^4$ and assign $S^i_u = \mathscr{P}_A(S^{i-1}_u, T^{i-1}_u)$ and $T^i_u = \mathscr{P}_B(S^{i-1}_u, T^{i-1}_u)$ respectively.
4: For every $v \notin \mathscr{F}$ and every leaf $u \in \mathscr{C}(v)$ assign $S^i_u = S^{i-1}_u$ and $T^i_u = T^{i-1}_u$.

In the rest of the proof we first analyze the correctness probability of the protocol above (the key lemma is Lemma 8.7) and then total communication (Lemma 8.10). The proof of Theorem 2.8 is completed by observing that the protocol can be executed in $O(r)$ rounds.

Lemma 8.7. *After stage i for every leaf $u \in \mathscr{T}$ it holds that $S^i_u = T^i_u$ with probability at least $1 - 1/(\operatorname{ilog}^{r-i-1} k)^4$, taken over all the randomness of the protocol.*

Proof. If u is in the subtree of a node v, which is not *failed* at level i then we know that $S_v = T_v$ and thus $S_u = T_u$ for each $u \in \mathscr{C}(v)$ with probability at least $1 - 1/(\operatorname{ilog}^{r-i-1} k)^4$ by the guarantee of the EQUALITY(S_v, T_v) test. Otherwise, u is in the subtree of a failed node v at level i. In this case the claim follows because we run BASIC-INTERSECTION protocol for this leaf with success probability at least $1 - 1/(\operatorname{ilog}^{r-i-1} k)^4$. □

We call a node $v \in L_i$ *correct* if after stage i it holds that $S_v^i = T_v^i$.

Corollary 8.8. *Every node $v \in L_i$ is correct with probability at least $1 - 1/(\text{ilog}^{r-i-1} k)^3$. In particular, the root of the tree is correct with probability at least $1 - 1/k^3$.*

Proof. From Lemma 8.7 applied to the level i it follows that after the execution of stage i for every leaf $u \in \mathscr{C}(v)$ it holds that $S_u^i = T_u^i$ with probability at least $1 - 1/(\text{ilog}^{r-i-1} k)^4$. Hence, by a union bound over all $\text{ilog}^{r-i} k$ such leaves with probability at least

$$1 - \text{ilog}^{r-i} k/(\text{ilog}^{r-i-1} k)^4 \geq 1 - 1/(\text{ilog}^{r-i-1} k)^3$$

we have $S_v^i = T_v^i$. □

The correctness proof of the protocol now follows from Corollary 8.8 together with the following invariant applied to the root of the tree after round $r-1$.

Proposition 8.9. *If for a node $v \in \mathscr{T}$ Alice and Bob assign S_v^i and T_v^i to it respectively then if $S_v^i = T_v^i$ then $S_v^i = T_v^i = S_v \cap T_v$.*

Proof. Note that this invariant is maintained by BASIC-INTERSECTION (Corollary 8.4). During the execution of the protocol the sets S_v' and T_v' only change when we apply BASIC-INTERSECTION to the leaves in $\mathscr{T}$. Clearly, if the invariant is maintained for all leaves then it is also maintained for all internal nodes as well. □

Now we analyze the total communication in the protocol. For a leaf $u \in \mathscr{T}$ let n_u denote the expected number of times the BASIC-INTERSECTION protocol was run on the sets assigned to u.

Lemma 8.10. *For every leaf $u \in \mathscr{T}$ it holds that $\mathbb{E}[n_u] = O(1)$.*

Proof. For a leaf u let's denote it's unique predecessor in level i as $p_i(u)$. Formally, $p_i(u) = v$ if and only if $v \in L_i$ and u is in the subtree of v. We can express $\mathbb{E}[n_u]$ as:

$$\begin{aligned}
\mathbb{E}[n_u] &= \sum_{i=0}^{r-1} \Pr[p_i(u) \text{ is failed}] \cdot (4\,\text{ilog}^{r-i} k) \\
&\leq \sum_{i=0}^{r-1} d_i \cdot \Pr[v \text{ is an incorrect child of } p_i(u)]\,(4\,\text{ilog}^{r-i} k), \\
&\leq \sum_{i=0}^{r-1} \frac{\text{ilog}^{r-i} k}{\text{ilog}^{r-i+1} k} \cdot \frac{1}{(\text{ilog}^{r-i} k)^3} \cdot (4\,\text{ilog}^{r-i} k) = O(1)
\end{aligned}$$

where the first inequality holds by a union bound and the second by Corollary 8.8. □

The total expected communication in the protocol can be expressed as the sum of the total communication for EQUALITY and BASIC-INTERSECTION. The total communication for EQUALITY is:

$$\begin{aligned}
&\sum_{i=0}^{r-1} |L_i|(4\,\text{ilog}^{r-i} k) \\
&= O(k\,\text{ilog}^r k) + \sum_{i=1}^{r-1} (k/\,\text{ilog}^{r-i} k) \cdot (4\,\text{ilog}^{r-i} k) \\
&= O(k\,\text{ilog}^r k) + O(rk) \\
&= O(k\,\text{ilog}^r k).
\end{aligned}$$

The expected total communication for BASIC-INTERSECTION is by Lemma 8.3 equal to:

$$\mathbb{E}\left[\sum_{i=1}^{k}(|S_i|+|T_i|)\log(|S_i|+|T_i|)\cdot n_i\right] = \sum_{i=1}^{k}\mathbb{E}\left[(|S_i|+|T_i|)\log(|S_i|+|T_i|)\right]\mathbb{E}[n_i],$$

where the equality follows from the independence of the random variables. Because for every i we have $\mathbb{E}[n_i] = O(1)$ by Lemma 8.10, to complete the proof it is sufficient to show that $\mathbb{E}[(|S_i|+|T_i|)\log(|S_i|+|T_i|)] = O(1)$ and thus the total communication for BASIC-INTERSECTION is $O(k)$. We have $\mathbb{E}[(|S_i|+|T_i|)\log(|S_i|+|T_i|)] \leq \mathbb{E}[(|S_i|+|T_i|)^2]$, where the right-hand side is constant by the same argument as used to bound each term in (25). Finally, the bound on the number of rounds of communication follows from the fact the communication in each of the r stages for the EQUALITY tests can be done in parallel in two rounds (Fact 8.5). After in four more rounds we can perform all BASIC-INTERSECTION protocols in parallel (Lemma 8.3). This gives $6r$ rounds of communication. □

9 Multi-Party Set Intersection in the Message Passing Model

In the multi-party case we have m players, each holding a set $S_i \subseteq [n]$ such that $|S_i| \leq k$. The goal of the parties is to output a set $S = \bigcap_{i=1}^{m} S_i$. We allow arbitrary communication between the parties (i.e. any player i can send a message to any player j). In each round of the protocol the parties first perform some local computation and then can exchange messages. This is known as the message passing model (see e.g. [7]). We consider two optimization goals: minimizing the total communication (or equivalently average communication per player) and minimizing the worst-case communication per player. In both cases we keep the number of rounds as small as possible.

First, observe that we can amplify the success probability of the two-party protocol in Theorem 2.8 to be $1-1/2^k$ while keeping the expected total communication $O(k\log^r k)$ and only incurring a penalty in the number of rounds: the protocol will have expected $O(r)$ rounds instead of worst-case $6r$ rounds. This follows by repeating the protocol if it hasn't succeeded. The latter condition can be checked by exchanging k-bit equality checks after the protocol terminates. With a total of $O(1)$ expected repetitions this gives expected $O(r)$ number of rounds and success probability which is only limited by the equality checks and is thus $1-1/2^k$ by Fact 8.5.

Using this observation we obtain a protocol with the following guarantee for the average-case multi-party setting.

Corollary 9.1. *(Average-case) For every $r > 0$ there exists a protocol for m-party Set Intersection in the message passing model with expected average communication per player $O(k\log^r k)$, expected number of rounds $O\left(r\cdot\max(1,\frac{\log m}{k})\right)$ and error probability $1-1/2^k$.*

Proof. First, the set of m players is partitioned into groups of size at most 2^k. Consider one such group, which consists of players holding sets $S_1,\ldots,S_{2^k}$. The player holding S_1 is chosen as a coordinator. Within the group all players execute the modified version of the two-party protocol described above with the coordinator, who computes sets $T_i = S_1 \cap S_i$ for each $2 \leq i \leq 2^k$. This step is repeated until the coordinator succeeds in verifying that $\bigcap_{i=2}^{2^k} T_i = \bigcap_{i=1}^{2^k} S_i$ with probability at least $1-1/2^k$. This is done by using a $2k$-bit equality check with each of the players. By Fact 8.5 the equality check succeeds with probability $1-1/2^{2k}$ and hence by a union bound over the 2^k players in the group the desired success probability follows. Once all $m' = \lceil m/2^k \rceil$ coordinators succeed in verifying their sets the protocol is executed recursively among them for their respective sets.

The number of active players decreases exponentially between the levels and thus the total communication is dominated by the first level. The first level has average complexity $O(k \operatorname{ilog}^r k)$ per player and expected $O(r)$ rounds using the same reasoning as for the case of two-parties discussed above. The total number of levels of recursion is $\max(1, \log_{2^k} m) = \max\{1, k^{-1} \log m\}$, which gives the claimed bound on the total number of rounds. □

Taking $r = \log^* k$ in Corollary 9.1 we get average communication $O(k)$ per player, which matches the lower bounds of [47, 7] who show that average communication $\Omega(k)$ is necessary for solving SET INTERSECTION and SET DISJOINTNESS in the message passing model.

In the protocol from Corollary 9.1 every coordinator has to perform $O(2^k k \operatorname{ilog}^r k)$ communication per level. By increasing the number of rounds we can amortize this cost among the players.

Corollary 9.2. *(Worst-case) For every $r > 0$ there exists a multi-party protocol in the message passing model with worst-case communication $O\left(k^2 \operatorname{ilog}^r k \cdot \max\left(1, \frac{\log m}{k}\right)\right)$ per player, expected number of rounds $O\left(rk \max(1, \frac{\log m}{k})\right)$ and error probability $1 - 1/2^k$.*

Proof. The protocol is executed recursively in $\max\left(1, \frac{\log m}{k}\right)$ levels and in each level the players are assigned to groups of size at most 2^k as in Corollary 9.1. Consider one such group. Instead of using a coordinator in each level the players are assigned to the leaves of a complete binary tree of depth k. They run the two-party protocol recursively in pairs. This gives expected number of rounds $O(rk)$ per level and the bound on the number of rounds follows. When the two-party protocol is executed for the top two nodes in the tree (the children of the root) the parties also perform a k-bit equality check in order to certify the correctness of the result with probability $1 - 1/2^k$. If this check fails then the entire computation in the tree is repeated, which gives $O(1)$ repetitions in expectation using the same reasoning as before. Finally, adding up over all nodes on a path of length k the worst-case communication per level is $O(k^2 \operatorname{ilog}^r k)$ which gives the bound on the desired worst-case communication per player. □

10 Concluding Remarks

We have gained new insight into the complexity of EQUALITY, one of the cornerstones of the theory of communication complexity. To do so, it was important to consider the *expected* communication cost of a protocol on a *fixed* input, and to limit the amount of interaction that our players can use. It was also important to treat 1-inputs (i.e., equal pairs) and 0-inputs separately. Our results have applications to other important communication problems, namely DISJOINTNESS and PRIVATE-INTERSECTION.

The upper bounds in Section 7.1 show that our OREQ and k-DISJ lower bounds are not absolutely improvable: they are already tight in small-error settings. One drawback in our direct-sum approach is that the error requirements in our OREQ and k-DISJ lower bounds needs to be similar to the (small) error for EQUALITY protocols. On the other hand, the Sağlam–Tardos approach [49], which directly attacks OREQ, overcomes this to obtain the same communication lower bound even under constant error. This raises the interesting theoretical question of whether a direct-sum approach can be strengthened to "boost" the error.

In recent work on direct sum questions in communication complexity, there has been some exciting progress on a related matter. Molinaro, Woodruff, and Yaroslavtsev [42] show how to obtain constant-error direct sum theorems from small-error hardness of the underlying problem. Unfortunately, their technique depends crucially on the k-fold direct sum problem's output being a k-tuple consisting of the solutions to *all* of the k independent instances of the underlying problem. In our setting, these k bits are combined into a single bit by an OR operation, which gives out much less information, causing their technique to

fail. Showing similar results for problems with a single-bit output is a challenging open problem whose resolution ought to yield even more insights about communication complexity.

Our results for PRIVATE-INTERSECTION raise two main open questions. First, it is open whether the number of rounds in our two-party protocol (Theorem 2.8) can be reduced from $6r$ to r while preserving the total communication of $O(k \operatorname{ilog}^r k)$. This would match results of [49] for k-DISJ. The second open question is to design better multi-party protocols than those that we obtain in Section 9. Our multi-party protocols are obtained by using our two-party protocol as a black-box and we expect that it might be possible to obtain better results by analyzing the problem directly.

References

[1] Farid Ablayev. Lower bounds for one-way probabilistic communication complexity and their application to space complexity. *Theoretical Computer Science*, 175(2):139–159, 1996.

[2] Anil Ada, Arkadev Chattopadhyay, Stephen A. Cook, Lila Fontes, Michal Koucký, and Toniann Pitassi. The hardness of being private. *ACM Transactions on Computation Theory*, 6(1):1:1–1:24, March 2014.

[3] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999. Preliminary version in *Proceedings of the 28th Annual ACM Symposium on the Theory of Computing*, pages 20–29, 1996.

[4] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences*, 68(4):702–732, 2004.

[5] Boaz Barak, Mark Braverman, Xi Chen, and Anup Rao. How to compress interactive communication. *SIAM Journal on Computing*, 42(3):1327–1363, 2013. Preliminary version in *Proceedings of the 41st Annual ACM Symposium on the Theory of Computing*, pages 67–76, 2010.

[6] Mark Braverman. Interactive information complexity. In *Proceedings of the 44th Annual ACM Symposium on the Theory of Computing*, pages 505–524, 2012.

[7] Mark Braverman, Faith Ellen, Rotem Oshman, Toniann Pitassi, and Vinod Vaikuntanathan. A tight bound for set disjointness in the message-passing model. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science*, pages 668–677, 2013.

[8] Mark Braverman, Ankit Garg, Denis Pankratov, and Omri Weinstein. From information to exact communication. In *Proceedings of the 45th Annual ACM Symposium on the Theory of Computing*, pages 151–160, 2013.

[9] Mark Braverman and Ankur Moitra. An information complexity approach to extended formulations. In *Proceedings of the 45th Annual ACM Symposium on the Theory of Computing*, pages 161–170, 2013.

[10] Mark Braverman and Anup Rao. Information equals amortized communication. In *Proceedings of the 52nd Annual IEEE Symposium on Foundations of Computer Science*, pages 748–757, 2011.

[11] Mark Braverman, Anup Rao, Omri Weinstein, and Amir Yehudayoff. Direct product via round-preserving compression. In *Proceedings of the 40th International Colloquium on Automata, Languages and Programming*, pages 232–243, 2013.

[12] Joshua Brody, Amit Chakrabarti, and Ranganath Kondapally. Certifying equality with limited interaction. Technical Report TR12-153, ECCC, 2012.

[13] Harry Buhrman, David Garcıa-Soriano, Arie Matsliah, and Ronald de Wolf. The non-adaptive query complexity of testing k-parities. *Chicago Journal OF Theoretical Computer Science*, 6:1–11, 2013.

[14] Amit Chakrabarti, Graham Cormode, Ranganath Kondapally, and Andrew McGregor. Information cost tradeoffs for augmented index and streaming language recognition. *SIAM Journal on Computing*, 42(1):61–83, 2013. Preliminary version in *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, pages 387–396, 2010.

[15] Amit Chakrabarti, Subhash Khot, and Xiaodong Sun. Near-optimal lower bounds on the multi-party communication complexity of set disjointness. In *Proceedings of the 18th Annual IEEE Conference on Computational Complexity*, pages 107–117, 2003.

[16] Amit Chakrabarti and Ranganath Kondapally. Everywhere-tight information cost tradeoffs for augmented index. In *Proceedings of the 15th International Workshop on Randomization and Approximation Techniques in Computer Science*, pages 448–459, 2011.

[17] Amit Chakrabarti, Yaoyun Shi, Anthony Wirth, and Andrew C. Yao. Informational complexity and the direct sum problem for simultaneous message complexity. In *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science*, pages 270–278, 2001.

[18] Amit Chakrabarti and Anna Shubina. Nearly private information retrieval. In *Proceedings of the 32nd International Symposium on Mathematical Foundations of Computer Science*, volume 4708 of *Lecture Notes in Computer Science*, pages 383–393, 2007.

[19] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2006.

[20] Anirban Dasgupta, Ravi Kumar, and D. Sivakumar. Sparse and lopsided set disjointness via information theory. In *Proceedings of the 16th International Workshop on Randomization and Approximation Techniques in Computer Science*, volume 7409, pages 517–528, 2012.

[21] Mayur Datar and S. Muthukrishnan. Estimating rarity and similarity over data stream windows. In *Proceedings of the 10th Annual European Symposium on Algorithms*, pages 323–334, 2002.

[22] Ronald Fagin, Moni Naor, and Peter Winkler. Comparing information without leaking it. *Communications of the ACM*, 39(5):77–85, 1996.

[23] Tomas Feder, Eyal Kushilevitz, Moni Naor, and Noam Nisan. Amortized communication complexity. *SIAM Journal on Computing*, 24(4):736–750, 1995. Preliminary version in *Proceedings of the 32nd Annual IEEE Symposium on Foundations of Computer Science*, pages 239–248, 1991.

[24] Michael L. Fredman, János Komlós, and Endre Szemerédi. Storing a sparse table with $O(1)$ worst case access time. *Journal of the ACM*, 31(3):538–544, 1984. Preliminary version in *Proceedings of the 23rd Annual IEEE Symposium on Foundations of Computer Science*, pages 165–169, 1982.

[25] Michael J. Freedman, Kobbi Nissim, and Benny Pinkas. Efficient private matching and set intersection. In *Advances in Cryptology-EUROCRYPT 2004*, pages 1–19, 2004.

[26] Rusins Freivalds. Probabilistic machines can use less running time. In *IFIP Congress*, pages 839–842, 1977.

[27] Andre Gronemeier. Asymptotically optimal lower bounds on the NIH-multi-party information complexity of the AND-function and disjointness. In *Proceedings of the 26th International Symposium on Theoretical Aspects of Computer Science*, pages 505–516, 2009.

[28] Prahladh Harsha, Rahul Jain, David McAllester, and Jaikumar Radhakrishnan. The communication complexity of correlation. In *Proceedings of the 22nd Annual IEEE Conference on Computational Complexity*, pages 10–23, 2007.

[29] Johan Hastad and Avi Wigderson. The randomized communication complexity of set disjointness. *Theory of Computing*, 3(1):211–219, 2007.

[30] Rahul Jain. New strong direct product results in communication complexity. *Electronic Colloquium on Computational Complexity (ECCC)*, 18:24, 2011.

[31] Rahul Jain, Attila Pereszlényi, and Penghui Yao. A direct product theorem for the two-party bounded-round public-coin communication complexity. In *Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science*, pages 167–176, 2012.

[32] Rahul Jain, Pranab Sen, and Jaikumar Radhakrishnan. Optimal direct sum and privacy trade-off results for quantum and classical communication complexity. *CoRR*, abs/0807.1267, 2008.

[33] Bala Kalyanasundaram and Georg Schnitger. The probabilistic communication complexity of set intersection. *SIAM Journal on Discrete Mathematics*, 5(4):547–557, 1992.

[34] Daniel M. Kane, Jelani Nelson, and David P. Woodruff. On the exact space complexity of sketching and streaming small norms. In *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1161–1178, 2010.

[35] Iordanis Kerenidis and Ronald de Wolf. Exponential lower bound for 2-query locally decodable codes. *Journal of Computer and System Sciences*, 69(3):395–420, 2004. Preliminary version in *Proceedings of the 35th Annual ACM Symposium on the Theory of Computing*, pages 106–115, 2003.

[36] Hartmut Klauck. On quantum and approximate privacy. *Theory of Computing Systems*, 37(1):221–246, 2004. Preliminary version in *Proceedings of the 19th International Symposium on Theoretical Aspects of Computer Science*, pages 335-346, 2002.

[37] Eyal Kushilevitz and Noam Nisan. *Communication Complexity*. Cambridge University Press, Cambridge, 1997.

[38] Eyal Kushilevitz and Enav Weinreb. The communication complexity of set-disjointness with small sets and 0-1 intersection. In *Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science*, pages 63–72, 2009.

[39] Frédéric Magniez, Claire Mathieu, and Ashwin Nayak. Recognizing well-parenthesized expressions in the streaming model. In *Proceedings of the 41st Annual ACM Symposium on the Theory of Computing*, pages 261–270, 2010.

[40] Kurt Mehlhorn and Erik M. Schmidt. Las Vegas is better than determinism in VLSI and distributed computing (extended abstract). In *Proceedings of the 14th Annual ACM Symposium on the Theory of Computing*, pages 330–337, 1982.

[41] Peter Bro Miltersen, Noam Nisan, Shmuel Safra, and Avi Wigderson. On data structures and asymmetric communication complexity. *Journal of Computer and System Sciences*, 57(1):37–49, 1998. Preliminary version in *Proceedings of the 27th Annual ACM Symposium on the Theory of Computing*, pages 103–111, 1995.

[42] Marco Molinaro, David Woodruff, and Grigory Yaroslavtsev. Beating the direct sum theorem in communication complexity with implications for sketching. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2013.

[43] Moni Naor and Benny Pinkas. Oblivious polynomial evaluation. *SIAM Journal on Computing*, 35(5):1254–1281, 2006.

[44] Ashwin Nayak. Optimal lower bounds for quantum automata and random access codes. In *Proceedings of the 40th Annual IEEE Symposium on Foundations of Computer Science*, pages 124–133, 1999.

[45] Ilan Newman. Private vs. common random bits in communication complexity. *Information Processing Letters*, 39(2):67–71, 1991.

[46] Mihai Pătraşcu. Unifying the landscape of cell-probe lower bounds. *SIAM Journal on Computing*, 40(3):827–847, 2011.

[47] Jeff M. Phillips, Elad Verbin, and Qin Zhang. Lower bounds for number-in-hand multiparty communication complexity, made easy. In *Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 486–501, 2012.

[48] Alexander Razborov. On the distributional complexity of disjointness. *Theoretical Computer Science*, 106(2):385–390, 1992. Preliminary version in *Proceedings of the 17th International Colloquium on Automata, Languages and Programming*, pages 249–253, 1990.

[49] Mert Saglam and Gábor Tardos. On the communication complexity of sparse set disjointness and exists-equal problems. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science*, pages 678–687, 2013.

[50] Andrew C. Yao. Probabilistic computations: Towards a unified measure of complexity. In *Proceedings of the 18th Annual IEEE Symposium on Foundations of Computer Science*, pages 222–227, 1977.

[51] Andrew C. Yao. Some complexity questions related to distributive computing. In *Proceedings of the 11th Annual ACM Symposium on the Theory of Computing*, pages 209–213, 1979.